nature

VERY PERSONAL GENOMICS

A father tries to understand his daughter's DNA

CLIMATE CHANGE
IPCC after the Nobel

GENETIC VARIATION
'HapMap3' raises the bar

NATURE.COM
Focus on genetics

nature

www.nature.com/nature

Vol 449 | Issue no. 7164 | 18 October 2007

Rising to the climate challenge

The award of a Nobel prize to an advisory body in the science of climate change rightly reflects the organization's many virtues, and should spur it on in its mission to assess and address global warming.

The Intergovernmental Panel on Climate Change (IPCC) is not a household name but it deserves to be. Its important and altruistic work during the past two decades fully merits this year's Nobel Peace Prize, which it shares with the former US vice-president Al Gore (see page 766).

The current understanding of anticipated climate change and its effect on ecosystems and societies, uncertainties and all, is not anecdotal. Rather, it is articulated explicitly as a consensus view of a worldwide community of researchers. Too few politicians and members of the public appreciate this. And although not every individual scientist involved will fully agree with each sentence and each probability estimate in the IPCC's reports, few if any will seriously question that what the IPCC delivers is as good a chunk of scientific advice on climate change as anyone could hope to get.

In its latest, fourth assessment, a synthesis of which will be released next month, the IPCC has compiled the strongest evidence so far that the current warming trend is the increasingly dangerous result of human activity. This is an apolitical statement. Taking the appropriate political steps is the responsibility of the countries that, in Bali in December, will continue negotiating a follow-up agreement to the 1997 Kyoto Protocol on climate change.

The challenge of climate change urgently demands a coherent political response. But no matter whether and how soon an agreement on stabilizing greenhouse-gas emissions at safe levels might be reached, the IPCC should continue its successful work in informing policy-makers on how the science is developing, and on further climate signals and trends as they are uncovered.

Even if policy-makers can mitigate the effects of climate change, adaptation to them will still be a necessity. But how much, and when, and where? Perhaps the biggest challenge for the climate-research community over the next five years will be to use new scenarios of mitigation and adaptation to generate predictions about how climate change will affect specific regions. Climate modellers are already gearing up to address the problem of predicting regional climate change out to 2030 or so. The results of their increasingly detailed projections are likely to become the core of a fifth IPCC assessment in five or six years.

Although much of the IPCC's strength lies in the very scale and thoroughness of its assessments, reports of more restricted scope may also be desirable in the shorter term. There are a number of specific threats that merit deeper assessment, such as the risks of rising sea level and retreating sea-ice in the Arctic, and the effects of feedback loops on the climate system.

"The challenge of climate change urgently demands a coherent political response."

There are other important issues that would also benefit from being reviewed promptly by the IPCC. The economic costs of mitigation scenarios, including doing nothing, should be addressed, as should the vexed but persistent debates surrounding engineered attempts to influence the climate system.

Many climate scientists would like to move away from an IPCC process in which three independent working groups that investigate science, impacts and mitigation, respectively, work almost entirely independently of each other. But the established process is difficult to avoid in drawing up a full-scale assessment, and any suggestion of a merger should be resisted: assessing mitigation is best kept separate from assessing science if only to support the objectivity of the latter. More focused studies can involve greater interaction between, say, climate modellers, impact researchers, economists and coastal engineers.

Climate science — both measurement and modelling — will develop rapidly over the next few years, and alarm may grow with further insights. But a fifth assessment by the IPCC should not be rushed. This Nobel peace laureate is an organization whose strengths include an understanding that, however urgent the challenge, robust scientific advice, like science itself, needs patience. ■

Do-it-yourself science

How much involvement can patient advocates have in genetics?

Some of the hype that accompanied the first publications of the human genome sequence in 2001 may have worn off. But these are still heady times for genomics, as demonstrated this week by the release of a greatly enhanced haplotype map or HapMap, which describes the most common forms of human genetic variation (see page 851).

The map builds on an earlier version published in 2005. It may, for example, shed some light on aspects of the genome that help to account for certain differences between people of different geographical origins (see page 762). There have been plenty of other research findings this year that demonstrate the power of genomics to deliver clues that could yield better medicine, including studies based on the HapMap that have uncovered lists of multiple genes that may be associated with the risk of developing specific diseases.

But there remain relatively few examples where this has led to better treatment options for patients and doctors. The difficulties of selecting relevant gene and protein markers, and then developing them into marketable tests that doctors will use, remain formidable (see page 770). And for patients, doctors and even some geneticists, there is growing frustration at the lack of clarity in some research

findings, the difficulty in discerning which findings are of medical value, and the slow pace at which fuller knowledge of the links between genetics and disease is actually providing better diagnosis and treatment options.

On page 772 of this issue, *Nature* tells the story of Hugh Rienhoff, a trained geneticist and biotechnology entrepreneur, whose daughter was born with a collection of congenital defects. He has taken it upon himself to try to find out what the genetic cause might be — actually buying lab equipment and having her genes sequenced himself. He has even posted information about her condition, his theories as to what's causing it, and parts of her genetic sequence on the Internet.

Given the sharply falling costs of equipment and the wealth of information that is publicly available, we are getting to the point at which almost anyone with access to the Internet can do this. If that sounds a little scary, then perhaps it ought to. Scientists and patient advocates have always enjoyed a delicate relationship. Researchers are not prone to welcome what they may see as the intrusion of the public in the laboratory. And there is every chance that some people in Rienhoff's position will waste money pursuing dead ends. On the other hand, as more people begin to take an interest in rare or undiscovered disorders, more useful information is likely to be unearthed about both their genetics and their treatment.

But this means that clinical geneticists will have to revise the professional and ethical framework for collaborating with patients and their advocates, to help ensure that the information from the public provides clarity and not confusion. Some scientists are already thinking about how best to organize such information. On page 783, for example, Steven Brenner of the University of California, Berkeley, proposes a 'genome commons' to aggregate the accumulated knowledge on human genetic and phenotypic diversity.

"Scientists and patient advocates have always enjoyed a delicate relationship."

At the same time, members of the public who choose to embrace a do-it-yourself approach to science need to be aware that they should not abandon existing, rational treatment options. And they should know that the fruits of their labours will rarely include the cast-iron answers that they may be seeking. For, as is so often the case in science, the most likely result of their efforts will be yet more unanswered questions for others to probe. ■

Criteria creep

The politically motivated extension of a US stem-cell registry makes no scientific sense.

A steady lament from American biologists is that the human embryonic stem-cell lines that they can work on using federal research money are too old, and too few in number. But researchers will draw scant comfort from a White House executive order, issued on 20 June, that could sharply increase the number of cell lines on the Human Embryonic Stem Cell Registry — the list of cell lines that can be studied with support from the National Institutes of Health (NIH).

The executive order was issued by President Bush on the same day that he vetoed legislation that would have permitted funding to be used on research with additional embryonic stem-cell lines. It will replace the word 'embyronic' with 'pluripotent' in the registry's name, and thus add adult stem-cell lines to the registry — even though these are already eligible for federal funds. This political sleight of hand seems intended to increase the number of cell lines listed without adding new lines of embryonic cells.

For this to happen, senior NIH officials must now waste time trying to establish unarguable criteria that will affirm a cell line as pluripotent (see www.nature.com/stemcells). By the end of this month, they are expected to release an application form for researchers wishing to add new, non-embryonic lines to the list.

Pluripotency — which basically means that a cell can grow into any sort of body cell — is commonly evaluated in mouse cells by mixing candidate cells into mouse embryos and observing their subsequent development. But equivalent experiments cannot be ethically conducted with human cells, leaving no robust method for confirming their pluripotency. Indeed, pluripotency has yet to be formally proved in human cells *in vitro* — including in embryonic cells. There is also currently no way to prove that cells derived from embryonic and non-embryonic sources have equivalent capabilities to generate the specialized cells that could be useful in drug discovery and cell therapy.

The executive order calls for the thorough cataloguing of stem cells derived in what it calls "ethically responsible ways" — meaning, in the White House's parlance, that they are derived without creating, harming or destroying an embryo. The order further calls for the prioritization of new grants to study these lines. But no additional money is being allocated for this work, meaning that it can only proceed at the expense of other research supported by the NIH.

Flexible cells from non-embryonic sources do offer exciting possibilities: perhaps adult human cells can be reprogrammed, and cells from the testis and amniotic fluid can be coaxed into an array of functioning tissues. If such cells can be derived from individual patients with diseases, these non-embryonic sources could be of great value. But this value is more likely to be unleashed if they are studied alongside embryonic stem cells, rather than in their place.

On 9 August 2001, when Bush first announced his restriction of federal research funding to embryonic stem-cell lines already derived by that date, his officials suggested that researchers would be able to work with about 60 lines. But the true number has turned out to be about 20, of which only a dozen are commonly used. NIH director Elias Zerhouni told a Senate committee back in March that the range of embryonic stem cells currently available to US researchers is insufficient, and is hampering scientific innovation and biomedical research.

The NIH has been obliged by law to come up with a plan for implementing the executive order. It will no doubt make the best of a difficult situation, and come up with some criteria for pluripotency. It is regrettable that one of the world's leading research agencies should be required to make avowedly scientific distinctions along lines drawn up to suit the administration's political requirements. ■

RESEARCH HIGHLIGHTS

Parallel protection

Proc. R. Soc. B doi:10.1098/rspb.2007.1039 (2007)

To win the game of concealment, it's often best to use the tools at hand. Several desert spiders from around the world hide by attaching sand to their bodies. Using scanning electron microscopy on moults of spiders from Africa, South America and the United States, undergraduate Rebecca Duncan and her colleagues from Lewis & Clark College in Portland, Oregon, compared two unrelated spider genera that independently evolved this ability.

The team found that both have long, thin, flexible 'hairlettes' on the bristles that cover their bodies. The researchers suggest that intermolecular forces make sand stick (see inset, right) and keep the spiders camouflaged.

The almost indistinguishable methods for adhesion in the two genera show the power of evolution to produce similar adaptations in similar environments.

K. CRAMER, MONMOUTH COLL., ILLINOIS

NEUROBIOLOGY

Faulty transmission

Neuron **56,** 58–65 (2007)

Researchers have found that a protein implicated in the autism-spectrum disorder 'Rett syndrome' regulates the formation of certain neuronal connections.

Loss of the protein MeCP2 causes Rett syndrome, but gene duplications that double protein levels also produce autism-like characteristics and seizures. A likely cause of these disorders is an imbalance between excitation and inhibition in the brain. Hsiao-Tuan Chao, Huda Zoghbi and Christian Rosenmund at Baylor College of Medicine in Houston, Texas, found that when mice lack MeCP2, neurons that transmit the amino acid glutamate in the hippocampus showed 46% less transmission. And mice that produced twice the normal level of MeCP2 exhibited twofold higher neuronal transmission.

These changes were primarily due to alterations in the number of connections — known as synapses — between neurons, the researchers found. The results suggest that MeCP2 regulates synapse formation during early development.

MECHANICS

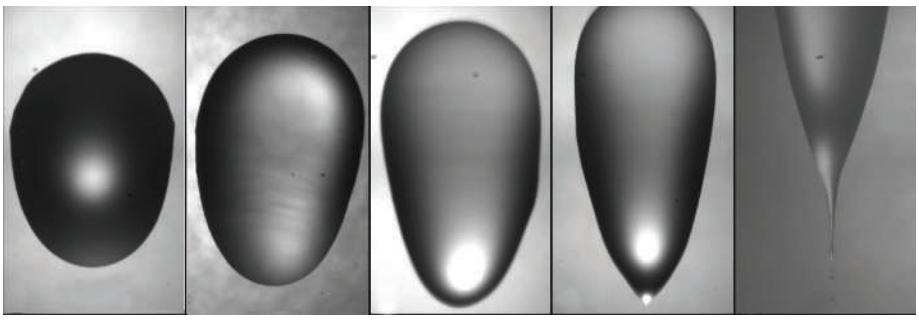
A defiant droplet

Phys. Rev. Lett. **99,** 144501 (2007)

How can water travel uphill? Simple, say Philippe Brunet and colleagues from the University of Bristol, UK: place a droplet on a sloping plate and shake the plate up and down.

As the plate descends, the droplet gets taller, with a larger contact angle on the upper than the lower edge (pictured, below). This tends to push it up the slope. The effect should cancel out over a complete oscillation cycle, but it doesn't owing to a combination of factors: fore- and aft-symmetry breaking due to the slope, and nonlinear friction between the droplet and the surface.

Above a certain vibration amplitude and frequency all drops will climb. This might be used to transport liquids in microfluidic networks.



AM. PHYS. SOC.

ASTRONOMY

Superlative supernova

Astrophys. J. Lett. **668,** L99–L102 (2007)

Astronomers have found the most luminous supernova yet.

Named 2005ap, it lies 4.7 billion light years away and burst with a peak magnitude roughly eight times as bright as the Milky Way. Discovered just before this peak in March 2005, the supernova faded after three weeks. It was about twice as bright as the previous bar-setter, discovered by the leader of the current team, Robert Quimby, a postdoctoral researcher at the California Institute of Technology in Pasadena and leader of the Texas Supernova Search.

The authors speculate that this supernova's record-setting luminosity may have come from an exploding red giant's shock wave hitting and lighting up a shell of material around the star.

MICROBIOLOGY

Life on the rocks

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0708183104 (2007)

Exactly how some microorganisms live in glacial ice has been something of a mystery. Previous research identified two habitats in which they could obtain water and energy — the surface of trapped mineral grains and liquid veins at ice boundaries. Some life is tuned to these niches, but microbes unable to survive in them have nevertheless been found in glacial ice, meaning another

habitat is probably present.

Buford Price and Robert Rohde, at the University of California, Berkeley, may have identified this missing habitat. They calculated that enough molecules such as carbon dioxide, oxygen, nitrogen and methane can diffuse through ice to sustain life.

By scanning ice cores with laser fluorimeters they detected protein spikes, some of which were indicative of single isolated cells, in just such habitats.

ASTROCHEMISTRY

Salty stars

Astrophys. J. **668,** L131–L134 (2007)

Researchers in the United States have found a dash of the unexpected in oxygen-rich stars. Lucy Ziurys and her colleagues at the University of Arizona in Tucson used the Submillimeter Telescope on Mount Graham and the 12 Meter Telescope on Kitt Peak, both operated by the Arizona Radio Observatory, to observe two red-giant stars that have shells dominated by oxygen. By analysing the recorded spectra, the team determined that the shells contain NaCl, which has previously been observed only in carbon-rich red giants.

The findings suggest that oxygen-rich stars, like their carbon-rich cousins, may be home to the complex types of chemistry that create molecular precursors to life.

BIOCHEMISTRY

Keeping the 'code'

Cell **131,** 58–69 (2007)

Certain chemical changes, or marks, made to the histone proteins around which DNA wraps seem to tell the cell whether or not that DNA should be transcribed.

Teams led by Matthias Mann at the Max Planck Institute for Biochemistry in Martinsried, Germany, and Marc Timmers at the University Medical Centre Utrecht in the Netherlands looked for proteins that bind to one chemical mark — trimethylation of lysine 4 on the histone H3. This mark is usually associated with transcriptional activity, and they found that a component of the transcription factor TFIID bound it tightly.

Dimethylation of a nearby arginine residue inhibited this binding, and other specific marks strengthened it, lending credence to the hypothesis that a combinatorial 'histone code' determines how cells read their DNA.

PLANT ECOLOGY

Grass attack

J. Ecol. doi:10.1111/j.1365-2745.2007.01307.x (2007)

Looking for signs of biological warfare past, Carolyn Malmstrom of Michigan State University in East Lansing and her colleagues delved into herbarium specimens at two University of California sites and extracted some of the oldest plant-virus RNA ever recovered.

Although ecological theory generally says that invasive species are successful outside their home ranges because they are freed from the pathogens that evolved to plague them, Malmstrom and colleagues suspect that a historical takeover of California grasslands by Eurasian grasses succeeded in part because the invaders brought viruses with them that affected the natives or changed the dynamics of an existing virus population.

They extracted barley yellow dwarf virus RNA from several specimens, including a 1917 invasive wild oat, proving that the virus was present at the time of invasion.

VISION

A scaffold in new light

EYE OF SCIENCE/SPL

Cell **131,** 80–92 (2007)

The fruitfly protein INAD had long been considered to be a scaffolding protein, organizing important visual signalling proteins that attach to it. But recent research suggests that INAD directly regulates visual perception.

Rama Ranganathan, of the University of Texas Southwestern Medical Center in Dallas, and colleagues show that, in response to light, one of five structural 'PDZ' domains of INAD transiently switches from a reduced to an oxidized state, distorting INAD's ability to bind to other molecules. This seems crucial to visually mediated reflex behaviours and for terminating visual responses.

Many scaffolding proteins contain PDZ domains, which could undergo similar conformational changes to that of INAD. Thus, rather than support components, these might serve as control centres for other signalling molecules.

Correction

The Research Highlight 'Volcanic paintings' (*Nature* **449,** 510; 2007) wrongly named Joseph Mallord William Turner as John Mallord William Turner.

JOURNAL CLUB

Andre Geim
University of Manchester, UK

Imploding atoms have softened this experimentalist's teasing views on theoretical physics.

As an experimentalist, I instinctively dislike theory papers. Too many of them seem to be written for the sole purpose of showing off an integral larger than a competitor's, or to present multiple theories just in case one idea proves right and so is hailed as visionary. I feel even less warmly towards theories that are nigh on impossible to check, such as the supposed precursor to a theory of everything, string theory.

But speaking seriously, even the most obscure predictions can turn out to be spectacularly relevant.

In our lab we have been studying graphene, a material that comprises a single layer of carbon atoms arranged similarly to chicken wire. Because electrons in this material mimic ultra-relativistic particles, it should be possible to observe in their behaviour century-long-predicted phenomena such as the Klein paradox (which concerns how highly energetic electrons tunnel through supposedly impenetrable barriers) and *zitterbewegung* (jittery movements of relativistic wave-packets).

Several recent theory papers on the physics preprint server arXiv predict another coup for graphene (see A. V. Shytov *et al.* arXiv:0708.0837; 2007).

According to relativistic quantum theory, atoms containing more than 170 protons cannot exist, because electrons around nuclei with such a large charge would fall into the centre. Nuclear physicists have not come close to creating atoms heavy enough to test this prediction. But the recent theory papers suggest that it should be relatively easy to observe the effect in graphene. This is because electrons in this material interact much more strongly than they do in atoms, so should fall down on charged impurities (standing in for nuclei) rather routinely.

This makes me wonder: could we design condensed-matter systems to test the supposedly non-testable predictions of string theory too?

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

French agency head resigns in cancer row

The resignation on 8 October of Christian Bréchot, director-general of the French medical research agency, INSERM, is the latest and highest-profile fallout of a bitter conflict about new technologies for screening cancer cells. The dispute pits Bréchot's wife, Patrizia Paterlini-Bréchot, an INSERM oncologist, against the current management of Metagenex, a company the couple founded in 2001. The clash has triggered lawsuits and investigations by France's top ethical and health authorities.

The pair created the firm to commercialize a filtration technique they developed called ISET (isolation by size of epithelial tumour cells). ISET can detect circulating tumour cells (CTCs) in the blood, picking out individual tumour cells from billions of blood cells (G. Vona *et al. Am. J. Pathol.* **156,** 57–63; 2000). Detection of CTCs is a hot area of cancer research. Combined with research on genetic markers, it promises insight into disease progression and metastasis; it could lead to better tailored treatments or, more speculatively, earlier diagnosis.

Promising technique

Klaus Pantel, chairman of the Institute of Tumor Biology at the University of Hamburg in Germany and an expert in CTCs, says he is a "very strong protagonist" of CTC technology: "I want it to go into the clinic." But he warns against excessive hype for any one system, saying, "it's a complex field". All of the existing systems for analysing CTCs, including contenders from Immunicon of Huntingdon Valley, Pennsylvania, and Veridex, a subsidiary of Johnson & Johnson based in New Jersey, as well as ISET, have their "pros and cons", he adds. "We mustn't push too early, and create too high expectations."

"This whole story is bizarre."
— Claude Kordon

Whether high or not, Metagenex says that its legitimate expectations have been thwarted by the company's founders. On 13 July it filed a lawsuit claiming €43.2 million (US$61.3 million) in damages against Bréchot, Paterlini-Bréchot and INSERM's technology transfer arm, INSERM Transfer. The suit claims that Bréchot and Paterlini-Bréchot "voluntarily paralysed" the company by blocking its access to exclusive licences held by INSERM, the Assistance Publique-Hôpitaux de Paris, and the University of Paris V, promised under a rider to an agreement between the parties.

David Znaty, the company's director-general, claims that this was part of a wider campaign by Paterlini-Bréchot, in particular, to damage the company and eventually force him out. The terms of his contract require him to sell his shares to her in such circumstances. The suit claims Paterlini-Bréchot's motivation was to regain control of the company, a result that the sale of Znaty's shares would provide. On 28 September, Paterlini-Bréchot filed a countersuit against Znaty for the same sum.

Bréchot, who like his wife denies any wrong doing, says that his resignation was in part so that he can have his hands free to fight the case; he also talks of pursuing new career challenges. He vehemently denies the company's accusation that, as the chief official at INSERM, he faced a conflict of interest in decisions about licensing after disagreements arose between his wife and Metagenex's management. He is not a shareholder himself, having divested his shares when he took up his INSERM position. The shares are now held by the couple's children.

The conflict between the company's founders and managers began shortly after Znaty, an engineer and entrepreneur trained at the Massachusetts Institute of Technology, was invited by Paterlini-Bréchot to join the company in April 2006 when it was at risk of bankruptcy. But it got worse after a €2.5-million investment by AXA, a financial services company, and Banexi, a venture-capital firm, three months later. This reduced the Bréchot family's shareholding from 83% to 44%.

Paterlini-Bréchot alleges that Metagenex went on to commercialize the ISET technology without fully disclosing its experimental nature, that it broke an agreement to pursue multicentre trials to clinically validate applications for the follow-up of cancer patients, and that it illegally used the test for the commercial diagnosis of cancer. Znaty flatly rejects the allegations as a "smokescreen". His suit in turn accuses Paterlini-Bréchot of various breaches.

Patrizia Paterlini-Bréchot (above) is suing the company she and her husband (left) founded.

T. BOUËT; E. FEFERBERG/AFP/GETTY

The legal battle is only one side of the dispute, which is also being fought in the arena of medical ethics and public opinion. Paterlini-Bréchot has personally solicited investigations into the Metagenex affair by the National Consultative Ethics Committee for Health and Life Sciences (CCNE) and INSERM's own ethical committee, billing the Metagenex case as being central to the ethics of developing diagnostics for severe diseases, such as cancer. Reports from both bodies backed the principle that such tests need rigorous multicentre trials to validate their sensitivity and specificity, and to assess any correlation with disease progression. As a result, the government has decided to clarify relevant legislation and to close up some loopholes.

An ethical dilemma?

Bréchot says that the results of these investigations vindicate the July 2006 decision by the patent holders to block Metagenex's licences until it was clear that the test had been validated clinically. Given that the ethical and legislative situation has now been made clearer, he says, there should no longer be any obstacle to Metagenex having access to the patents. That is perhaps putting a brave face on the matter; Bréchot's resignation was prompted by a letter to INSERM from the research and health ministries on 21 September. The letter said that in

CONFERENCE BLOG
Catch up with the meeting of the American Society for Reproductive Medicine.
http://blogs.nature.com/news/blog

view of their own ongoing investigations into the Metagenex affair they thought it "in the best public interest" to allow the company to exploit the patents.

Yvon Cayre, an INSERM haematologist at the Hôpital Robert Debré in Paris who replaced Paterlini-Bréchot as Metagenex's scientific consultant in January 2007, says the ethical debate has nothing to do with the company. He says he has shifted the company away from Paterlini-Bréchot's focus on cancer diagnostics to refinement of the machine designed to let researchers and doctors "use CTCs as a non-invasive way to survey treatments and specially targeted therapies in patients with cancer".

Claude Kordon, a retired INSERM researcher who coordinated the CCNE panel on the Metagenex affair, says he was surprised that Paterlini-Bréchot brought the ethical issues before the committee. It is "not an ethical problem. It is about a conflict between parties who signed contracts, and did not agree on the interpretation of the contract," says Kordon. "This whole story is bizarre."

The court cases are not the only ongoing investigation. The IGAS, the powerful 'police' of the French health system, is due to release the results of yet another investigation shortly. It has sent a preliminary version to the parties to allow their comments. The draft version concludes that the reports by the two ethics committees "cannot be used to justify the refusal to sign the rider", and that Paterlini-Bréchot began to pose ethical questions only as a result of conflict in the company. She denies this absolutely. ■

Declan Butler

Particle collider is on schedule...just

Rumours of construction delays at the world's largest particle accelerator have exaggerated the size of the problem, according to the project's head. "There have been no show stoppers," wrote Robert Aymar, director-general of CERN, the particle physics lab near Geneva, Switzerland, in the 8 October issue of the lab's *CERN Bulletin*. "We can all look forward to the LHC producing its first physics in 2008."

His reassuring announcement came after gossip on physics blogs of new problems that could set the lab's Large Hadron Collider (LHC) even further behind its already delayed start date. But for all the reassurance, the LHC schedule remains tight, says project leader Lyn Evans. Relatively small mishaps could push the opening back beyond July 2008, when the LHC is supposed to start doing physics.

The machine is a CHF10-billion (US$8.4-billion) accelerator designed to slam protons into each other at energies of up to 14 teraelectronvolts — 7 times the current record. Detectors will comb through the debris from the collisions for evidence of the Higgs mechanism, which is believed to endow all other particles with mass, and for signs of physics beyond the 'standard model', the current theoretical framework of particle physics.

LHC construction has already faced a number of setbacks and delays. For example, in March, a support holding a superconducting magnet in one section of the machine failed, leaving engineers scrambling for a fix (see *Nature* doi:10.1038/news070402-3; 2007).

The latest round of speculation was sparked in September when engineers found that several pieces of tubing between magnets had crumpled as the machine was being prepared for servicing. Just six out of 424 sections collapsed, according to Raymond Veness, who leads the LHC's vacuum engineering team. But the failures blocked the beam line, along which particles will eventually travel. "This is a potentially nasty problem," Veness says.

For the time being, engineers have improvised a solution. Using compressed air, they are firing a small plastic ball equipped with a radio-transmitter through the line. By detecting where the ball comes to a halt, they have been able to spot collapsed sections in a matter of minutes. But changing the defective sections, a half-day task, adds to an already full maintenance schedule. "Everyone is stretched in all directions," adds Veness.

"I don't think anyone sees it as an insurmountable problem," says Peter Limon, a high-energy physicist at Fermilab in Batavia, Illinois, the only other lab currently working with very high-energy protons. But whether the July date can be met will depend on how the magnet systems behave in additional tests this winter, says Evans. A magnet or other component failure in a section of the ring cooled to liquid helium temperatures for testing could set things back by months because of the time needed to warm the section up for repairs and cool it back down.

"The next three months are going to be pretty critical," says Evans. "If something unforeseen comes up between now and then, it will slip. There's no doubt." ■

Geoff Brumfiel

D. PARKER/SPL

The Large Hadron Collider, a 27-kilometre accelerator, is due to start operation in July 2008.

SPECIAL REPORT

So similar, yet so different

Tiny pieces of the genome can already explain many human characteristics. **Erika Check Hayden** looks at what they might reveal in the future.

In his 2000 State of the Union Address, President Bill Clinton chose to emphasize something he had recently heard from a genome researcher: that humans are all, irrespective of race, 99.9% the same genetically. "Modern science," he told his country's legislators, "has confirmed what ancient faiths have always taught: the most important fact of life is our common humanity." Seven years on, and four years after the final publication of the sequences from the Human Genome Project, new technologies and larger data sets are allowing genome biologists to answer a conundrum embodied in that unity-inspiring percentage: if our DNA is so similar, why do we seem different in so many ways?

The answer, in part, is that the genome is not as uniform as Clinton was led to believe; nor is it nearly as sedate, stable and homogeneous as scientists used to think. It's less a 'Book of Life', more a wiki; many of its ho-hum elements don't change, but some really interesting bits are constantly revised.

"Maybe 99% of our genome behaves in a nice, predictable way," says Gilean McVean, a statistical geneticist at the University of Oxford, UK. "But it's become clear that there is this pool of errant variants that are responsible for a lot of the dynamism in our genome, and we don't understand its consequences for disease risk or normal variation."

Over the past year, two large studies[1,2] have found evidence that many people carry around lots of large chunks of DNA that are deleted, copied, flipped or otherwise rearranged in other people. The findings confirm earlier studies that hinted at this type of 'structural' variation but were not large enough to command assent[3]. A study of the genome of sequencing pioneer Craig Venter also found much more variation from reference sequences than expected[4].

The larger analyses estimate that such variable regions could make up more than 10% of the genome, vindicating scientists, such as Evan Eichler of the University of Washington in Seattle, who have long argued that structural variation is a major source of diversity. Scientists are still investigating how much it contributes to differences between populations. But it is already clearly linked to some differences between individuals that can be correlated with behaviour or environment. For example, a study published in September reported[5] that evolution has driven a starch-digestion gene to duplicate itself in people with traditionally starch-heavy diets.

"We're getting away from this 0.1% figure that has been in our minds ever since the draft human genome sequence came out," says Hunt Willard, head of the Institute for Genome Sciences and Policy at Duke University Medical Center in Durham, North Carolina. "We're now looking at maybe half a per cent of content that is unique to individual genomes." The actual variation is thus lower than the extent of the variable regions, but larger than previously thought. "Maybe Eichler always had that number in his head," adds Willard, "but no one else did."

That said, Willard points out that population geneticists such as Luca Cavalli-Sforza and Richard Lewontin arrived at a similar figure in pioneering studies linking protein and gene diversity to the history of human populations. What's different today is the scale and type of data available.

The HapMap, a catalogue of the variation seen in 270 people from America, Japan, China and

H. IBRAHIM/CORBIS

All change: human genomes seem not to be as uniform as previously claimed.

West Africa, is a case in point. Today, McVean and a cast of hundreds publish a second-generation analysis of the HapMap (see page 851), the first phase of which was published in 2005 (ref. 6). The new, more thorough version finds surprisingly high diversity in single nucleotide polymorphisms (SNPs) — parts of the genome marked out by specific changes in a single DNA base pair.

The HapMap uses SNPs to identify chunks of DNA that tend to stay the same within populations. Researchers can then create an ordered list of SNPs, and thus DNA chunks, for each chromosome and choose one 'tag SNP' to stand in for the many SNPs that travel together on each chunk. But in the updated HapMap, 1% of the more than 3 million SNPs that have now been analysed cannot be grouped with their neighbours to mark identical chunks of DNA. These 'untaggable SNPs' reveal parts of the genome that vary greatly between people. "These untaggable SNPs are completely doing their own thing," McVean says. "It's not a high percentage of SNPs, but it's still a lot of them."

R. EDMONDS/AP

Bill Clinton: impressed by the revelatory power of DNA.

Branching out

Scientists are now obtaining DNA from seven more populations with African, Asian and European ancestry that could help explain the origin of the mystery SNPs. They are also discussing a massive new bout of sequencing in an international project involving Chinese, British and US funders that would use new technologies to sequence the genomes of 1,000 individuals. Along with the two individual genome sequences already released[4], these data will fuel a field that is set to explode over the next year: the hunt for genetic signatures that discriminate between smaller and smaller groups.

"The HapMap data can clearly tell you whether you are African or Chinese, but the question becomes, how far can you take that?" asks population geneticist Carlos Bustamante from Cornell University in Ithaca, New York. "Can you predict whether somebody comes from one village or another? We are going to see all kinds of stuff we would never have imagined was possible."

But does this just amount to expensive, and possibly divisive, genealogy? Pardis Sabeti at the Broad Institute in Cambridge, Massachusetts, thinks not. Today, she publishes an extensive study that uses the HapMap to identify specific genes linked to human diversity (see page 913). Over the past three years, Sabeti and other scientists have performed a series of studies finding evidence for 'positive selection' in chunks of DNA that differ between populations — indicating that genes are evolving differently in people from different parts of the world.

Sabeti now reports that she has pinpointed specific genes that seem to be responsible for some of the positive selection affecting these chunks. For instance, she found that variants of two genes linked to infection with the Lassa virus are favoured in West Africans.

Sabeti hopes that such studies will help guide scientists towards biological pathways involved in such regionally specific diseases. But her work also raises the sensitive issue of the biological meaning and relevance of race. Variants peculiar to Asian populations in another pair of genes — linked to hair, teeth and sweat glands — have no obvious links to disease. And there is the possibility that such population-specific variations might lead in uncomfortable directions.

In 2005, for instance, geneticist Bruce Lahn from the University of Chicago in Illinois suggested that two genes linked to brain size had evolved rapidly in groups that migrated out of Africa tens of thousands of years ago[7,8]. His results prompted criticism among fellow scientists, who felt that he didn't have the proper evidence to back such an incendiary claim. Sabeti notes with relief that Lahn's genes haven't turned up in any genome-wide scan so far — another sign that his conclusions were unfounded. Lahn says that true tests of his work are beyond the scope of these approaches, and that he is using other methods, including resequencing parts of the genome, to bolster his conclusions.

"This is a very delicate time, and a dangerous time, as people start to come up with things that the general public, or the media, or various groups might misinterpret," Sabeti says. "I like the fact that, so far, the evidence we find for natural selection in humans is only skin deep." ■

1. Redon, R. *et al. Nature* **444,** 444–454 (2006).
2. Stranger, B. E. *et al. Science* **315,** 848–853 (2007).
3. *Nature* **437,** 1084–1086 (2005).
4. *Nature* **447,** 358–359 (2007).
5. Perry, G. H. *et al. Nature Genet.* **39,** 1256–1260 (2007).
6. The International HapMap Consortium *Nature* **437,** 1299–1320 (2005).
7. Evans, P. D. *et al. Science* **309,** 1717–1720 (2005).
8. Mekel-Bobrov, N. *et al. Science* **309,** 1720–1722 (2005).

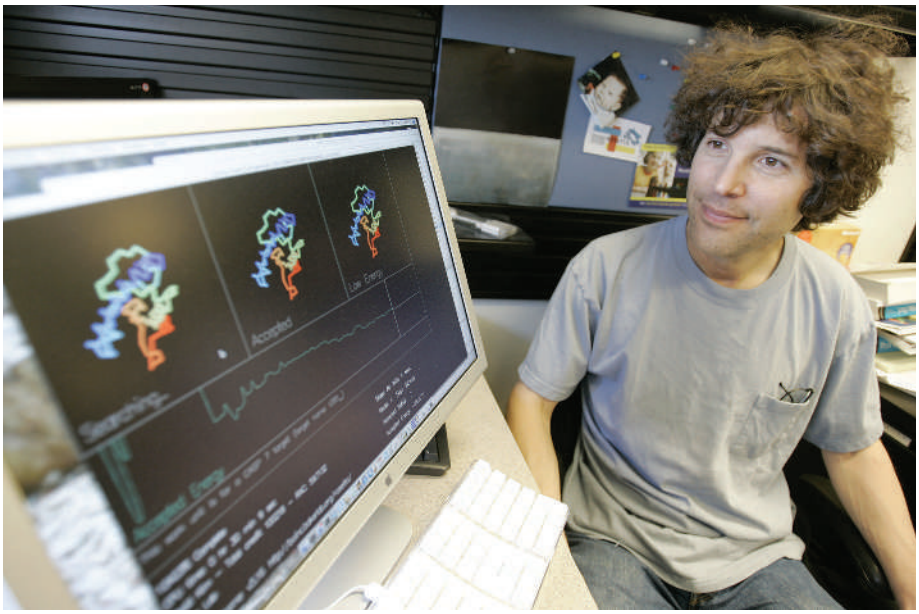
See http://tinyurl.com/3xw89t for an interview with Luca Cavalli-Sforza about his Italian Genome Project and geographic genetic diversity.
See also Editorial, page 755, and Books & Arts, page 785.

FIRST ASIAN GENOME SEQUENCED
Individual genomes get boost from Chinese effort.
www.nature.com/news

The shape of protein structures to come

By exploiting millions of hours of computing time donated by the users of 150,000 home computers, scientists have predicted the structure of a protein using just its sequence of amino acids. The project marks a significant advance in a field that's been flush with hope yet short on tangible results, experts say.

Determining the shape of a protein is normally a matter of firing X-ray beams at its crystalline form and measuring their diffraction, and protein chemists have long been sceptical of attempts to replace this practice with modelling or theory. "Modelling right now has a terrible name in the field," says Michael Levitt, a computational biologist at Stanford University. But in a *Nature* paper published online on 14 October, David Baker, a biochemist at the University of Washington in Seattle, and his colleagues report results that may do much to dispel that scepticism (B. Qian *et al. Nature* doi:10.1038/nature06249; 2007).



T. S. WARREN/AP

David Baker admires the handiwork of 150,000 computers.

A protein's shape — and therefore its activity — is determined by the precise way its constituent string of amino acids folds up. "If you think of the whole thing as like an utterly flexible snake, there are hundreds of degrees of freedom at every site," says Eleanor Dodson, a structural biologist at the University of York, UK. The final shape depends on the molecular interactions each amino acid has with its neighbours, with surrounding water molecules and with other amino acids that are a long distance away in terms of sequence, but which become close as a result of folding. It's a horrendous problem to model.

Rather than trying to solve the problem from first principles, Baker's technique combines information from the sum of what is already known about protein structures with the vast computing power available through the Berkeley Open Infrastructure for Network Computing. This software, developed at the University of California, Berkeley, allows people to contribute spare computing power on their desktops to scientific projects (most famously, the search for extraterrestrial intelligence, in the form of SETI@home); 150,000 volunteers used it to download a copy of the Baker lab's Rosetta@home program.

> **"If you think of the whole sequence as like an utterly flexible snake, there are hundreds of degrees of freedom at every site."**

Rosetta breaks a protein's sequence into short stretches that can be matched to identical stretches from proteins with known structures. These shapes offer many ways to sew the protein under study back together, and the program chooses those that minimize the free energy of the structure (a measure of its stability). By running the program over and over again on thousands of computers, the researchers lurched towards an ever more accurate protein model.

When fed the sequence of T0283, a 112-amino-acid protein from a bacterium, the network spat out several million structures after a million hours of computing time. Those millions were whittled down to five by further repeated computer analysis, and one of the structures was spot on, correlating with the structure as determined from its crystal.

Although the structure's precision was not on a par with high-resolution crystal models, it was good enough for researchers to think that the technique could simplify the process of obtaining X-ray structures in the future. To turn X-ray patterns into structures, researchers must produce patterns from crystals that have either been spiked with heavy-metal 'markers' or come with some indication of what the final structure will look like, for example from the shape of a related protein. Rosetta's structure for T0283 was good enough to function in this way. The program should now be able to provide such reference points for proteins for which there are already X-ray data, but which lack useful relatives and whose structures thus languish unsolved. "You will be able to solve a whole bunch of these structures rather quickly," says Adrian Roitberg, a protein modeller at the University of Florida in Gainesville.

There is still room for improvement, though, says Rhiju Das, a postdoc working on the Rosetta@home project and a co-author on Baker's paper. Each home computer works in isolation, he explains. If the program could be rewritten to run on the many parallel processors in a supercomputer, Rosetta might become considerably more powerful.

One pay-off of better structure prediction would be the prospect of custom-made proteins, says Baker, who uses Rosetta to hunt for sequences that correspond to desired structures. His lab and that of University of Washington biochemist Bill Schief are currently working on redesigning the gp120 protein of HIV to make a vaccine that could stimulate the immune system in a different way from the natural virus. The reshaped protein should elicit antibodies that attack the virus more effectively than antibodies created after infection.

The days when protein modellers thought they could make crystallization obsolete are long gone, Baker adds. But melding the two techniques could offer biologists insight into many more proteins — and faster. "If you really care about the structure of your protein, you should get some experimental data and combine it with modelling," he says. ■

Ewen Callaway

NATURE|Vol 449|18 October 2007

ON THE RECORD

"I have paid a certain fee to the word 'sustainability' to make it mean whatever I want it to."

Nobel laureate Murray Gell-Mann introduces his talk on 'Complexity and sustainability' at last week's Global Sustainability symposium in Potsdam, Germany.

SCORECARD

Bond gadgets
A Purdue University study has found that newly developed biometric sensors are no dirtier than your average doorknob.

Bond laboratories
UK defence contractor BAE Systems is being sued for accidentally destroying a business partner's laboratory while testing a new secret weapon.

3 GOOD REASONS...

To get to know your robot

RONALD GRANT ARCHIVE

1 Japanese researchers have developed a robot that gives professional-grade facial massages.

2 Another Japanese team has developed a 'green screen' mannequin robot that can take on the appearance of a fetching lady or muscled man.

3 Noticing the trend, a Dutch graduate student has successfully defended his PhD thesis predicting robot–human marriages by 2050.

Sources: Purdue University, Daily Telegraph, AP, Tech.co.uk, LiveScience.com

SIDELINES

RAVEENDRAN/AFP/GETTY

Jubilation: Rajendra Pachauri (left) and colleagues celebrate the IPCC's Nobel Peace Prize.

Climate change: a Nobel cause

"We had absolutely no clue, it's fantastic," says Stefan Rahmstorf, a climate modeller at the Potsdam Institute for Climate Impact Research in Germany. On 8–10 October, Rahmstorf, along with more than a dozen Nobel laureates and many others, was at a symposium in Potsdam discussing climate change and related problems under the rubric "Global sustainability: a Nobel cause". On 12 October he and they heard that the Nobel Peace Prize had been awarded to Al Gore, moviemaker and former US vice-president, and the Intergovernmental Panel on Climate Change (IPCC) — their cause had a Nobel prize of its own.

"Al Gore is a tireless fighter for the cause of climate who thoroughly deserves the prize for bringing this serious issue to the attention of politicians and the public," says Robert Watson, one of the Potsdam participants and a former IPCC chairman. "But I am incredibly pleasantly surprised that the IPCC will share the prize, which justly rewards the whole climate research community." Rajendra Pachauri, current chairman of the IPCC, also sees the prize as recognition of a global achievement. "I would like to pay tribute to the scientific community, who are the winners of this award," he says.

As climate change is ultimately a threat to peace and security, the Nobel Peace Prize is a totally appropriate reward for the IPCC's work, Watson adds. The prize has been given for environmental work before. In 2004 it went to Kenyan biologist and environmentalist Wangari Muta Maathai — who delivered a speech to the Potsdam meeting via live video from Nairobi — for initiating the 'Green Belt Movement' in Africa.

The IPCC has issued four assessment reports over the past two decades. Each is an extensive compendium of the available scientific literature on the science and effects of climate change, the result of work by thousands of authors and reviewers. "I can tell you a thing or two about sifting through piles and piles of reviews on evenings and weekends," says Rahmstorf, a lead author for the most recent assessment on the physical science behind climate change.

Discussions have recently begun on how to make information compiled by the IPCC more useful and relevant to policy-makers. Pachauri has drafted and disseminated notes listing several options, one being to issue shorter, more targeted and more frequent reports on special issues. Another possibility is to focus increasingly on the regional impacts of climate change. These and other options will be discussed at a series of meetings in the coming months. "We may have to modify the IPCC to make it more efficient," says Watson. "But the Nobel prize just reminds us that we need to continue."

Environmental advocates and many US scientists lauded Gore's efforts to bring climate science to a global audience, which began more than two decades ago and most recently took the form of the documentary film *An Inconvenient Truth*. Michael MacCracken, chief scientist for climate-change programmes at the Climate Institute in Washington DC, acknowledges that Gore was unable to move the United States towards action when he was vice-president under Bill Clinton in the 1990s, but says that since then Gore has been enormously suc-

cessful in spreading the word. "It's been really hard to capture public attention," he says, "and I think that is what Gore has really done."

Susan Solomon, a senior scientist with the National Oceanic and Atmospheric Administration and co-chair of the IPCC's first working group, calls the prize a "wonderful victory for science" and credits Gore with delivering the message to the public. "I love the movie. I really do," Solomon says. "I think his goal is to make people aware, and I think that is a very good goal. I don't think he has tried to promote a political agenda." Dan Schrag, a geochemist at Harvard University, while agreeing that the prize for Gore was well deserved, is not so sanguine about the politics. "The only concern I have is that Gore has helped make the climate issue a partisan issue in the United States, and that is not true in other parts of the world."

Predictably, the prize has increased speculation that Gore will make another run for the White House, a possibility that Gore has repeatedly played down but never entirely ruled out. ■

Quirin Schiermeier and Jeff Tollefson

See Editorial, page 755.

A. SCHMIDT/AFP/GETTY SUNDAY ALAMBA/AP/PA

HEALTH OFFICIALS FEAR NIGERIA POLIO SETBACK

Outbreak linked to vaccine may hit immunization rates.

www.nature.com/news

Happy birthday

This year's Nobel Prize in Chemistry was awarded to Gerhard Ertl (pictured) on 10 October, which just happened to be his seventy-first birthday. Ertl developed ways of applying spectroscopic techniques to tease out the mechanisms of reactions at surfaces, such as the Haber–Bosch synthesis of ammonia. He retired as director of the Max Planck Society's Fritz Haber Institute in Berlin in 2004. The prize committee said his work "provided the scientific basis of modern surface chemistry".

Chemists welcomed the award of this year's prize to a 'proper chemist', as several recent Nobels have gone to research that might in many eyes be seen more as biology. Some, though, have expressed surprise that the award was given to Ertl alone, and not shared with fellow surface chemist Gabor Somorjai of the University of California, Berkeley. When asked about this, Somorjai replied: "It's a very well deserved prize. Professor Ertl is a very good scientist, a good colleague and a good friend."

Daniel Cressey

For full coverage see http://tinyurl.com/26nyxh

Italian mafia accused of trafficking nuclear waste

Italy's anti-mafia squad has launched an official investigation into allegations of illegal trafficking and disposal of nuclear waste — as well as clandestine production of plutonium — by managers of the Italian National Agency for New Technologies, Energy and the Environment (ENEA).

Eight former employees of ENEA's Trisaia research centre in the southern town of Rotondella, and two alleged members of the 'Ndrangheta mafia, are under suspicion following a decade-long inquiry. Trisaia is now a multidisciplinary research centre, but in the 1970s and 1980s it specialized in nuclear waste processing and storage.

A mafia informer told the anti-mafia bureau in Potenza that an ENEA manager paid the 'Ndrangheta mafia to get rid of 600 drums of nuclear and toxic waste from Germany, France, Switzerland and the United States in 1987. He claimed that the mafia disposed of the radioactive material at unauthorized, non-secure sites in southern Italy, Somalia and in the Mediterranean Sea.

Investigators also suspect that the centre illegally produced plutonium during the 1980s, which the mafia allegedly sent to Iraq. ENEA denies all charges and says that the centre did not have the capacity to produce plutonium. "But we will collaborate fully with the investigations to dispel any suspicion of misconduct," says its president, Luigi Paganetto.

Scandinavian seals hit by deadly virus

An unknown virus killed more than 2,300 seals around the Scandinavian coast this summer, local scientists reported last week. The death toll, currently about 14% of the population, is likely to rise further, they say.

The virus attacks the seals' respiratory systems. They suffocate in their own mucus, and most die offshore. Scientists have recently seen breathing difficulties in some small dolphins in the area, suggesting the virus may also be infecting that species.

R. DIETZ/DMU, DENMARK

A mystery virus is killing Scandinavia's seals.

Tero Härkönen, a seal researcher at the Swedish Museum of Natural History in Kärna and his colleagues say the outbreak spread from the small Danish island of Anholt to the Skagerrak Strait, which flows between Denmark, Norway and Sweden — and then up to the Oslo Fjord.

A different virus attacked seals in Northern Europe in 1988 and 2002, wiping out about half of the population on each occasion. "The dynamics of the spreading of the two viruses are very similar," says Härkönen.

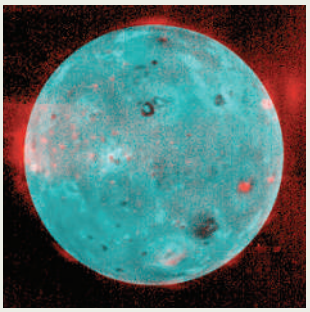
Virologists from the National Veterinary Institute in Uppsala, Sweden, are trying to identify the virus from samples taken from four dying seals.

Earth scientists aim to gauge US thirst

A US water census — the first in 25 years — could be on its way. The US Geological Survey has set a water audit as one of its decadal goals, noting that the issue will become more important as climate change alters water budgets.

Legislators have taken an initial step

Massive volcanic eruption seen on jovian moon



NASA/JOHNS HOPKINS UNIV./SOUTHWEST RES. INST.

Astronomers have witnessed a spectacular eruption on Io, the fiery moon of Jupiter. NASA's New Horizons probe snapped pictures (right) of a volcanic plume 350 kilometres high — 40 times the height of Mount Everest — when the probe passed by Jupiter in February and March on its way to Pluto.

Results from the fly-by were presented at the annual meeting of the American Astronomical Society's Division for Planetary Sciences on 9 October in Florida, and also appear as papers in the 12 October issue of *Science*.

Scientists were surprised to find that Jupiter's weather was abnormally quiescent. But Io — kneaded by Jupiter's gravity into constant volcanic upheaval — didn't disappoint. Scientists took 39 pictures of the Tvashtar volcano over 8 days, as an orange, sulphur-rich plume rocketed out at speeds up to 700 metres per second.

in authorizing the agency to carry out the census. On 4 October, US Senator Jeff Bingaman (Democrat, New Mexico) introduced a bill, the SECURE Water Act, that would expand the survey's groundwater and stream-flow monitoring programmes.

Equal pay for women in science is achievable

Aggressive academic management can correct pay disparities between male and female scientists, say researchers. Their study assesses the effects of intervention to equalize salaries at the University of Arizona's College of Medicine in Tucson between 2000 and 2004 (A. L. Wright *et al. J. Gen. Intern. Med.* 22, 1398–1402 ; 2007). By 2004, women with basic science doctorates, for instance, were paid 97.6% of the amount men were paid.

Lead author Anne Wright, the college's associate dean for faculty affairs, says the study was undertaken to gauge the success of administrative actions after an earlier analysis found women faculty members were paid about $13,000 (11%) less than men (A. L. Wright *et al. Acad. Med.* 78, 500–508; 2003).

The new study directly involved personnel records of about 400 faculty members, anthropologist Wright says, rather than using a percentage of staff who respond to a survey.

Cancer institute named after benefactor

The Massachusetts Institute of Technology (MIT) in Cambridge, will replace its Center for Cancer Research with the David H. Koch Institute for Integrative Cancer Research, it announced on 9 October.

Koch, a 67-year-old businessman and MIT alumnus, is battling cancer. He has committed $100 million to help build the facility, which should be completed in 2010 at a cost of $280 million. MIT bent its rules for Koch by committing to build the institute before it had the usual 80% of building costs.

BEST OF *NATURE*

Have you ever seen something in *Nature* — be it a research paper, news story or an editorial — that you thought deserved far more attention that it received? We value your opinion, so we've launched a website, 'Best of *Nature*', that allows readers to nominate, vote for and discuss content from *Nature*'s past. Visit http://spotlight.nature.com/bestofnature and tell us what we may have missed while compiling the 'History of the Journal *Nature*', a newly launched website which explores *Nature*'s history back to the very first issue in 1869.

BUSINESS

Missing the mark

Genetic tests to detect cancer are feasible. But with researchers drowning in a sea of biomarkers and little financial incentive to get the tests on the shelves, the idea is floundering. **Virginia Gewin** reports.

The genomic revolution brought in its wake the promise that it would be possible to detect —and arrest — the earliest signs of cancer. By tapping in to molecular biomarkers, initial signs of disease would be seen and effective treatments implemented. That promise remains resolutely unfulfilled.

In laboratories around the world, the discovery of candidate biomarkers continues apace. But the process of developing them into diagnostic tests is stalling. No early detection diagnostic test has so far been approved by regulatory bodies. Industry observers say a major problem is that there is no way to validate candidate biomarkers with the certainty that would merit their development into a marketable test. And the sheer number of candidates makes it hard to select which ones should be more fully developed.

> **"It's like the Wild West — anyone can stake claim to a biomarker."**
> — Sam Hanash

For their part, doctors want tests that identify the proper treatment course for individual patients. It is unclear whether they will adopt biomarker tests that serve only as diagnostics and don't point the way to specific therapies.

"Biomarkers are only as good as their ability to link to the treatment or pathology of the disease," says Harvey Pass, a clinician working on detection biomarkers for mesothelioma at New York University Medical Center.

Back in 2000, the US National Cancer Institute established the Early Detection Research Network (EDRN) in a bid to bridge the validation gap. The network sought to establish some of the infrastructure needed to test and validate candidate biomarkers. This is complicated by the fact that diagnostic tests often need to feature several markers to ensure that they are statistically significant.

On trial

In its draft assessment, the network reports movement on several fronts. Of the thousands of candidate biomarkers so far discovered, the EDRN is backing 120 in various stages of development. Among these, biomarkers for five cancers — mesothelioma, and liver, bladder, prostate and lung cancers — are in their third and final phase of development, and are being tested for effectiveness in large-scale human trials.

One of the EDRN's collaborators is Cangen Biotechnologies of Baltimore, Maryland, which hopes to put the first DNA-based early detection test for bladder cancer on the market. Cangen is confident that its test — a panel of 15 DNA segments — can identify tumours well ahead of existing diagnostics. But although detection may come a little earlier, a patient's primary treatment option remains the same: surgery. That means that the overall impact on patient survival rates may be relatively small.

For Eddy Agbo, Cangen's research director, linking the test to new drug therapies that could combat these early tumours is the next step forward. But in the absence of such therapies, there isn't a huge incentive for doctors and the health-care insurers that pay for most medical services in the United States to buy the tests.

Given its resources, the pharmaceutical industry would seem to be well placed to take candidate diagnostics tests through to approval. But rather than early detection tests, the industry seems more keen on using biomarkers to pinpoint patient suitability for treatment with particular drugs. There have been isolated successes for this approach, notably the targeting of the breast-cancer drug Herceptin, made by Genentech in South San Francisco, California. Herceptin targets patients who over-express a particular receptor and had sales of US$1.3 billion last year.

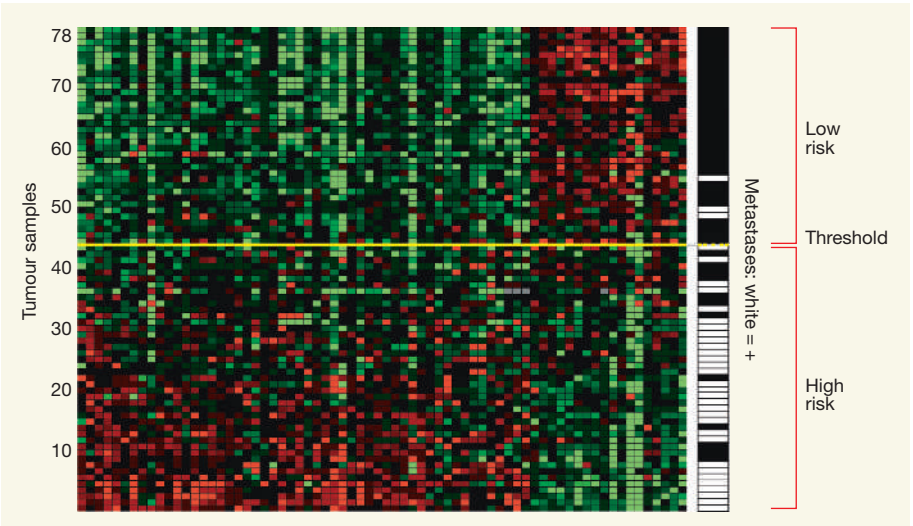
Everyone is looking for a dramatic demonstration of biomarker success, says Sam Hanash, a molecular biologist at the Fred Hutchinson Cancer Research Center in Seattle, Washington. The bottleneck isn't the lab-based discovery science, but the ability to efficiently weed through the overwhelming number of candidate genes, proteins and even microRNAs that could be clinically useful. "It's like the Wild West — anyone can stake claim to a biomarker," says Hanash.

Setting standards

The EDRN has done what it can to find some of the most promising markers and get them into trials. The network has made available cancerous and control reference tissue, serum and blood samples for lung, breast, ovarian, prostate and colon tumours. These can be used as the first step in validation by scientists claiming a biomarker. The EDRN has also established quality-control for validation testing, a standardized system for presenting data, and, with other groups at the National Cancer Institute, has established the beginnings of large-scale human biosample collection.

These steps are necessary, specialists say, but may not be sufficient to bring biomarker tests to market. The EDRN provides a valuable but incomplete resource, says Mike Gillette, a proteomics researcher at the Broad Institute in Cambridge, Massachusetts. Further validation steps require large-scale trials on broader populations, which are expensive and so temper industry interest.

In July, the US Food and Drug Administration (FDA) addressed industry concerns



L. VAN 'T VEER *ET AL. NATURE* **415**, 530–536; 2002/AGENDA

Genetic tests, such as this one for breast cancer, can help to tune treatments.

about the approval process by publishing draft guidelines for tests that use multiple biomarkers. Yet to be finalized, the regulations make an important distinction: class II tests, which are merely diagnostic or prognostic, can be validated using trials on tissue-bank samples. But class III tests, which are those that also indicate which therapies should be used, must undergo a more costly trial in which patients are tracked in the future.

The shallow pockets of most diagnostic testing companies can't support the multimillion-dollar costs of even the class II trials. "You need pushers and drivers to develop biomarkers," says Pat Price, a clinical researcher for Cancer Research UK at the University of Manchester, "and those are usually the drug companies." Price predicts that the drug industry will become steadily more interested in biomarkers as it comes to see its future in more segmented markets for drugs effective in particular genetic subpopulations. She says that, in general, the middle ground between biomarker discovery and the clinic is neglected by funding agencies and so is unattractive to researchers.

The EDRN is an exception — but it spends its $28-million budget validating only diagnostic or prognostic biomarkers. According to Pass in New York, collaboration with the network has been invaluable to develop early detection biomarkers for mesothelioma. He and others are now pushing for the EDRN to develop biomarkers for therapeutic prediction to complete the continuum of treatment.

Murray Robinson, a senior vice-president at AVEO Pharmaceuticals in Cambridge, Massachusetts, says that the success of Herceptin has made parts of the pharmaceutical industry sit up and take notice.

Larger companies, such as Roche of Switzerland, are vigorously exploring the co-development of biomarkers for particular diseases and drugs to treat them. That approach may have diminished the firm's interest in stand-alone tests developed by smaller companies. René Bernards, chief scientific officer of Agendia, a cancer-diagnostics company in Amsterdam, is disappointed by the lack of interest drug companies have shown in his firm's test. MammaPrint, which predicts the likelihood of breast-cancer recurrence, is the first multigene prognostic test to be approved by the FDA.

"If it's clear that the FDA is going to regulate this space of biomarkers, and ours is the only company to clear a multigene biomarker through the FDA, you would think a few large pharmaceutical companies would want to collaborate," he says. But Bernards is still waiting for someone to bite; and patients may wait many years yet for the true dawning of the age of personalized medicine. ■

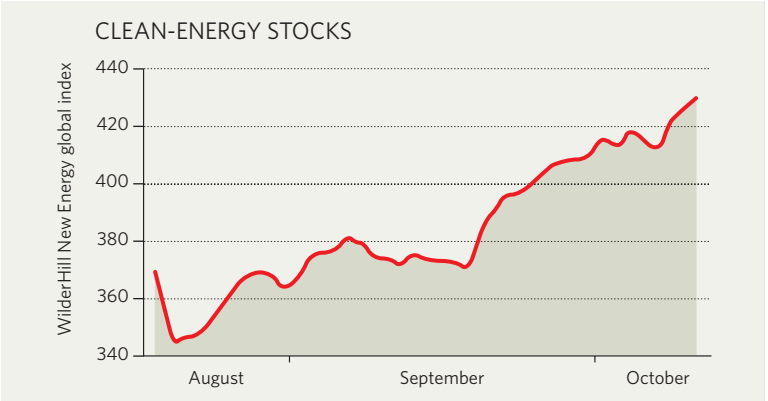
IN BRIEF

SCORING HIGH **Pharmaceuticals have surpassed computing and electronics hardware as the industry sector declaring the biggest total global investment in research and development (R&D), according to an annual scoreboard published by the European Commission. The scoreboard indicates that industrial R&D by the world's top 1,000 companies — thought to account for 85% of industrial R&D between them — surged ahead last year by 10%, to €372 billion (US$527 billion). US drugmaker Pfizer is now the biggest spender, investing €5.8 billion, whereas the largest-spending European company is DaimlerChrysler, at €5.2 billion.**

DRUG PRODUCTION **Ugandan president Yoweri Museveni has commissioned what his government says is the first factory in Africa that will produce full combinations of antiretroviral drugs for the treatment of AIDS. Quality Chemical Industries, a Ugandan drug importer, and Indian drug company Cipla will operate the plant, which is expected to start production early next year. The plant will also produce combinations of antimalarial drugs.**

AIMING LOW **Changes in UK rules on capital-gains tax could hit the emergence of London's Alternative Investment Market (AIM) as a popular global venue for listing science- and technology-based companies (see *Nature* 447, 367; 2007), some analysts say. The British finance minister, Alistair Darling, announced on 9 October that capital-gains tax would be levied at a flat rate of 18% from next April. Many investors on the AIM currently pay only 10%. But the announcement had no immediate effect on the prices of stocks listed on the AIM.**

MARKET WATCH



Stocks in clean-energy companies rebounded to reach all-time highs in the past two months, after dipping with the rest of the stock market earlier in the summer.

The WilderHill New Energy Global Innovation Index (symbol NEX on the American Stock Exchange) now stands at almost twice its level at the beginning of 2006 — reflecting the new-found tendency of mainstream investors to take 'alternative energy' stocks seriously.

A closer look at the index shows a mixed pattern, says Robert Wilder, whose California-based consultancy WilderShares co-compiles the index with New Energy Finance of London.

In particular, investors have realized that biofuels based on corn (maize) will not expand as rapidly as had been hoped, given growing concerns about rising food prices and the suitability of corn as a feeder fuel.

Biofuels stocks are near to their lowest point in the past year, Wilder notes, even as the index hit a new high on growing investor confidence in wind and solar-power companies. Wilder notes that many of the solar companies, in particular, are starting to perform like regular industrial firms in a rapidly expanded market, rather than as exotic start-ups. "The 'alternative' label is dropping off," he says.

There's clearly been a flood of money into stocks in these areas but, Wilder thinks, opportunities remain in sub-sectors that have so far been less fashionable, such as geothermal power and companies that specialize in energy efficiency. In the stock markets, at least, "efficiency hasn't been discovered yet", he says. ■

Colin Macilwain

HIS DAUGHTER'S DNA

Despite a training in clinical genetics, Hugh Rienhoff didn't know what was wrong with his daughter. So, as he tells **Brendan Maher**, he set about finding out.

Nearly four years ago, Hugh Rienhoff watched as his baby girl was pulled from a small incision in his wife's belly. It was their third child — the two boys had also been delivered by caesarean — and Rienhoff was there for all three births. But this child seemed different. He remembers her looking a little dark and sort of floppy, possibly attributable to the stress of delivery. Then he caught a glimpse of her feet, which were just a little longer than normal. For an instant, his training as a clinical geneticist kicked in. Could she have Marfan's syndrome?

In the joy of the moment the question vanished as quickly as it arose. "I didn't really think about anything from that point on medically, at least for that day," says Rienhoff. "I did all the usual things you do when you have a baby, which is cry and call my family." When the paediatrician handed his new daughter to Rienhoff, she offered some technical terms — nevus flammeus for a port-wine-stain birthmark down the middle of her face and arthrogryposis for the reluctance of her tiny fingers to extend all the way. Rienhoff had to write them down to remember them.

Although in the weeks and months after his daughter's birth the port-wine stain receded, it soon became clear to Rienhoff that she wasn't developing normally. Her fingers and toes wouldn't uncurl. More worryingly, in spite of ample feeding, the girl just wasn't gaining weight. Fleeting first impressions aside, she didn't have Marfan's, a disorder affecting maybe 1 in 5,000 births that arises from alterations in the gene for a protein called fibrillin-1. Skinny and birdlike with long fingers and flat feet, his daughter fit the physical characteristics of Marfan's fairly well, but she lacked some hallmark clinical criteria. In particular — thankfully — the typical defects in the cardiovascular system seemed to be absent, at least for now.

Rienhoff's daughter is one of the thousands of children born every year who have a congenital defect that resists satisfactory diagnosis. Such cases could be a known disorder that presents in an unusual way, or they could arise from a mutation rare enough not to have made it into textbooks or databases. Detailed genetic analyses are rarely undertaken on such cases unless a group of families with compelling commonalities can be found. Instead, these children are cared for as best can be.

But for Rienhoff that wouldn't do. Although he had largely left his practice, he had trained as a physician under Victor McKusick, the father of clinical genetics. Rienhoff knew genes, and he wanted to know his daughter's. For almost four years he has been trying to understand what makes her different at a molecular level, hoping that such knowledge could inform her care and treatment. He's quizzed experts, gone to meetings, and even set up gene-amplification equipment at home so that he can test his hypotheses with sequence data. He has also begun sharing the information he's found, telling his story on the Internet in the hope of helping others and of learning more. He may even have found a treatment that improves his daughter's condition.

****

As a medical student, intern, resident and research fellow, Rienhoff trained and worked at Johns Hopkins Hospital in Baltimore, Maryland, during the late 1970s and most of the 1980s. In 1992 he put clinical medicine and research largely behind him, leaving Johns Hopkins to become a partner with a Baltimore venture-capital firm, New Enterprise Associates. After years of helping biotech companies get off the ground, he decided to start one of his own. In 1998, he and his wife Lisa Hane moved to San Francisco where Rienhoff founded Kiva Genetics, later named DNA Sciences, a company aimed at developing a high-throughput sequencing platform for use in genetic discovery and diagnostics. He left his post there in 2001, and has continued to advise and found biotech ventures.

Confounding some expectations for a corporate type who has danced at the dizzying pace of start-ups for more than a decade, the 54-year-old Rienhoff is patient, thoughtful and soft spoken. He talks in lists as if every thought has been backed up by a careful tabulation of pros and cons. That's certainly how he has managed his daughter's care. With every new doctor she's seen, every test she's been given, there's been a meticulous calibration of the risks and of the benefits that she might receive.

One of his first carefully weighed decisions is one he remains adamant about: "I didn't want to be my daughter's doctor." (And in the con part of the table, he is quick to point out that he's not a paediatrician.) Even though he has begun to practise medicine again, Rienhoff has one relationship with her — as her father — and wants no other. He's happily drawn his own blood, but when he wanted to sequence his daughter's DNA, he took her to a phlebotomist. He couldn't bear to put her through pain.

****

Because of her arthrogryposis, the first doctors Rienhoff took his daughter to see were orthopaedists. They saw one ten days after her birth. "He was a thoughtful guy," remembers Rienhoff. "He said: 'This reminds me of

C. PICKENS

Beals', but it's not complete'." Beals' syndrome is a congenital disorder largely characterized by contracted joints, like those curling fingers and toes. Aside from that, the symptoms are quite similar to those of Marfan's, from which it was first distinguished some 35 years ago. The cause is similar to Marfan's, too: but in Beals' the mutation is in the gene for fibrillin-2 rather than fibrillin-1.

Being familiar with the genetics community has its perks. Rienhoff read some papers on the disorder and contacted the authors. One put him in touch with the eponymous Rodney Beals at Oregon Health and Science University in Portland. Beals, also an orthopaedist, responded that it didn't look like the syndrome he had described in 1971. Among other things, Beals' patients typically have their 'contractures' in larger joints than those of fingers and toes; but Rienhoff's daughter's knees and elbows were lax — indeed hyperextensible. Beals didn't think that he could help.

That said, Rienhoff knows all too well the difficulty in definitively ruling out disorders such as Marfan's and Beals'. They are genetically dominant, arising from a mutation in just one of the two copies people have for most genes. The mutant gene can be inherited from either parent, but that's not necessarily the case; sometimes a new mutation will crop up in sperm or egg. And because not all mutations in a gene will affect its expression, or the structure of its associated protein, in the same way, the symptoms associated with such a mutation can be quite different from the 'classic' form of the disease. Moreover, they may take years to manifest themselves. So Rienhoff's daughter may have a defect in fibrillin-1 or 2 that no one has ever seen before, and thus be a cryptic case of Marfan's or Beals'. But because there has never been enough evidence that his daughter has either of these diseases, she has not been sequenced for these genes, although she might be in the future.

"I think the most important thing that people could take away from this is that the process is not mysterious."
— Hugh Rienhoff

Rienhoff's first visit with a geneticist didn't provide much more clarity. The doctor suggested amyoplasia congenita, a diagnosis Rienhoff calls a relic, a "dustbin" for kids with various symptoms. And the collection of problems associated with this condition is so heterogeneous as to be useless. Like arthrogryposis, the term was merely a description of his daughter's symptoms. Thousands of children receive diagnoses like these, which don't shed light on what causes the problem or how it might be treated. "I couldn't go very far with that particular diagnosis. It was clear as we went forward that she had a syndrome," Rienhoff says. He believed her symptoms were related to each other and that they were probably caused by something specific in her genes.

Rienhoff's need for clarity was not purely intellectual. About five months after their daughter's birth, Rienhoff and Hane became very concerned about her failure to thrive. Although growing taller, she wasn't putting on weight. "She was just melting away," says Rienhoff. The gastrointestinal specialists they went to see advised them to stuff her with calories, but it didn't do any good. The doctors drew up a list of things that might be causing her problems — disorders of the metabolism, of the way nutrients were absorbed from gut and stomach, of the way that mitochondria in her cells produced energy. One possibility that arose was an unusual form of cystic fibrosis, but her symptoms looked quite unlike this condition.

Rienhoff thought that a mitochondrial disorder was a particularly plausible cause. The typical symptom is muscle weakness, which his daughter clearly had, but making a precise diagnosis is very tricky. Rienhoff dove into the literature and talked with the experts, quickly finding himself in what he calls a very messy field. "That really ate up a lot of time — eight or nine months," says Rienhoff.

As Rienhoff studied the murky world of the mitochondriacs, his daughter had her first birthday and took her first steps. She was developing — and, as a result, so was what could be said about her condition. When she stood up from a squatting position, she needed to brace her hands on her thighs. This behaviour, known as Gowers' sign, is common in children with muscle-wasting diseases such as Duchenne's muscular dystrophy. Sometimes, says Rienhoff, in a hard-to-determine diagnosis, you try to find a guiding principle. The inability to form muscle mass and tone, he says, "became the North Star of the case".

****

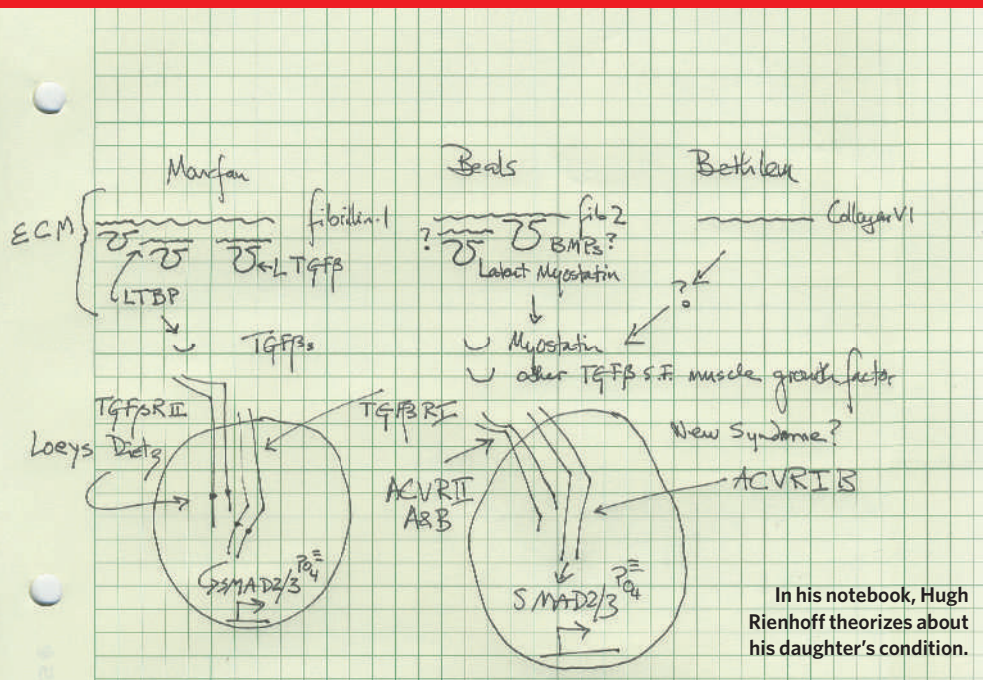
In the spring of 2005, Rienhoff and his daughter visited family and friends in Baltimore. He made an appointment with David Valle, a paediatric clinical geneticist whom he had met while working under McKusick. Valle, now director of the Institute of Genetic Medicine at Johns Hopkins, knows the limitations of his field as well as anybody. "Although there's been great progress in recent years, it still comes down to a careful [case] history, family history and physical exam, looking at the standard laboratory data and trying to put all these clinical features together to come up with some sort of diagnostic probability," he says. And even after a battery of standard genetic screening, "we're still left with maybe a third of patients who come in with morphological abnormalities for whom we're unable to make a diagnosis".

But when Valle was looking at Rienhoff's daughter with a couple of colleagues, something clicked. Her widely spaced eyes and marfanoid features, which are admittedly common in genetic disorders, looked strikingly similar to a syndrome that had just been defined. They asked the girl to open her mouth wide, and when they looked down her throat, they thought they'd cracked the case.

In January of that year Hal Dietz and Bart Loeys, both at Johns Hopkins at the time, had published a paper defining another condition similar to, and previously confounded with, Marfan's — Loeys–Dietz syndrome, to which they ascribed a related but distinct cause[1].

C. PICKENS

Hugh Rienhoff: father and basement DNA sequencer.



In his notebook, Hugh Rienhoff theorizes about his daughter's condition.

H. RIENHOFF

Fibrillin, the protein affected by mutations in Marfan's, is a structural component of the extracellular matrix, the protein netting that holds cells together. As a result it had long been assumed that the long, thin physical features and cardiovascular problems found in Marfan's were a result of the extracellular glue being structurally unsound. More recently, various lines of evidence, including research on Marfan's by Dietz and others, have suggested that the extracellular matrix does more than passively hold cells together; it mediates communication between them[2]. Fibrillin binds and sequesters the intercellular signalling molecules in the transforming growth factor-β (TGF-β) superfamily, which plays an important role in development. Fibrillin defects, it seems, free up the TGF-β signalling system with a range of effects: bones may grow extra long; vascular tissue may degrade.

Loeys and Dietz found that some people who seemed to have Marfan's harboured mutations not in the gene for fibrillin-1, but in the genes for two TGF-β receptors. Exactly how the mutations, which seem to disable the TGF-β receptors, have an activating effect on the pathway is an ongoing puzzle. But the result is a syndrome that, because it disrupts the same bodily system, is quite similar to that caused by the fibrillin-1 defects in Marfan's.

In addition to the molecular details, Loeys and Dietz had found three obvious bodily symptoms for their syndrome: the widely spaced eyes; a cleft in the palate and/or the uvula (the soft tissue that hangs down at the back of the throat); and severe structural defects in the arteries.

Strikingly — although it had never been noticed before, despite a great deal of medical and parental inspection — Rienhoff's daughter had a forked uvula. Valle and Loeys took blood samples to sequence the TGF-β receptor genes and suggested that the girl be given an echocardiogram as soon as possible. By chance, Rienhoff had scheduled one months earlier at the suggestion of David Clapham at the Children's Hospital in Boston, who suspected that her failure to thrive might be related to a heart defect. On the plane trip back Rienhoff read the Dietz and Loeys paper, which showed detailed pictures of the patients and their devastating aortic defects. His own heart sank.

****

"The problem is," Rienhoff said to me with a levelling look the first day I met him at a café in San Francisco, "she's amazingly cute." He smiled with the look of someone who knows that a desperate search for a diagnosis can sometimes end with a bad diagnosis. The average age of death for someone with Loeys–Dietz is 26. Over the year and a half of draining doctors' visits, Rienhoff and his daughter had developed a special bond. He has almost filled an entire notebook documenting his research on her case. But he has two more that are personal logs of his experiences with her, and casual observations of her and the funny things she says and does — he has similar notebooks for her brothers. One time, she asked him about a benign growth in the corner of his right eye. She calls him "Poppy", so his growth became a "poppy-oma".

On the Thursday after their return to the West Coast they went in for the heart exam. Rienhoff watched every moment of the echo, and her aorta came back "clean as a whistle" — a huge relief. The sequence data on the TGF-β receptors arrived a few weeks later; they showed none of the mutations Loeys and Dietz had identified for the syndrome. Dietz, whom Rienhoff went to visit the next year, says he wasn't completely surprised that the genetic testing came back negative. Rienhoff's daughter didn't have all the 'classic' symptoms associated with the syndrome.

It was reassuring that one dreadful diagnosis could largely be ruled out. But there was still the matter of what was actually going on. Inspired by Loeys, Dietz and Valle, Rienhoff found himself newly focused on the TGF-β signalling pathway. Maybe his daughter's disorder looked like Marfan's and Loeys–Dietz because related molecules were damaged. Rienhoff threw himself into the literature on TGF-β activation, once again guided by his North Star, his daughter's inability to build muscle.

There are dozens of different growth factors in the TGF-β superfamily and one of them, myostatin, is predominantly expressed in skeletal muscle. Defects in the gene for myostatin can result in overly muscled animals — notably the extraordinarily chunky Belgian blue cattle. In 2004, researchers in Germany and the United States described a young boy born to a former professional athlete[3]. He was remarkably muscular; at four-and-a-half he could hold two 3-kilogram weights at shoulder height with arms fully extended. Both his copies of the gene for myostatin were defective; his skeletal muscle was out of control.

Myostatin works through three activin receptors: ACVR1B, ACVR2A and ACVR2B. These look similar in sequence to the TGF-β receptors mutated in Loeys–Dietz. Rienhoff thought that a mutation in one of these specific receptors might explain why his daughter's skeletal muscle was so dramatically affected while her blood vessels were not. But as far as he knew no one had ever looked at them in relation to a disease. So he bought a used PCR machine, a microcentrifuge, some small-volume pipettes and a brand new gel box. All told, the equipment cost him about $2,000. With these simple tools and some sequence-specific DNA primers of his own design, he could pick the relevant genes out of his daughter's genome and amplify them enough for sequencing. Freezing the samples and packing the tiny tubes on ice, Rienhoff sent them off for sequencing at about $3.50 a pop. He prepared upwards of 200. If he was right, the data he got back would show a mutation in one of the genes for the activin receptors analogous to the mutations seen in Loeys–Dietz.

When he got the sequences, Rienhoff compared them to the human reference sequences in GenBank. In the gene encoding the ACVR1B receptor he found a variant. But it was a long way upstream of where he would have expected it to be, far from the active domain where many of the Loeys–Dietz mutations are found on the TGF-β receptor genes.

An obvious way to clear the mutation of any blame is for Rienhoff to sequence the copies of the gene in both his genome and his wife's. If one of them has the mutation too it is probably irrelevant — a harmless change, not one that explains the syndrome, because if it did the parent

Rienhoff, Hane and the children at home.

C. PICKENS

with the faulty copy would share the symptoms. Rienhoff says that he plans to sequence his and Hane's genes when he gets the time.

****

His notebooks are not the only record of Rienhoff's journey into his daughter's DNA. There's also mydaughtersdna.org, where he presents the clinical facts of his daughter's case both in layman's terms and also with the details required by a more medically astute audience. Although in some ways the website is a conduit for bloggish catharsis, Rienhoff sees it as an attempt to serve others with unidentified genetic disorders who are looking for answers. He hopes it will give others a chance to share their experiences, bring together parents and patient advocates, and perhaps even identify other patients with symptoms similar to his daughter's. The site got off to a slow start, but a few people have now begun posting their stories, including Rienhoff's colleague Clapham, who recounts the heartbreaking loss of his son, Ben, to an inexplicable neurological disorder.

With help from George Church, a Harvard Medical School professor with an extensive track record in new technologies for sequencing and synthesizing DNA, Rienhoff developed a sort of 'phenotype spreadsheet' on which to record his daughter's clinical history. The idea is that such data representations might someday allow a computer to search through his daughter's symptoms along with those of others with unidentified genetic disorders looking for clinical commonalities. Church plays down his contribution as just a seed of an idea that he thought worth testing, but says he is intrigued by Rienhoff's gumption: "I'm interested in cases of altruists who, rather than hiding from genetics, are using the opportunity to be sort of social activists, working to raise consciousness and maybe raise money for diseases affecting their family and friends."

Such activism is not new. Parents quite frequently become advocates for research on behalf of their children — when they are rich, famous, persistent, lucky or very well placed, they can make a difference. But with a sequencer and a website, Rienhoff has stepped over the threshold of personal genomics in a way set to catch the imagination. As sequencing gets ever easier and knowledge bases ever larger, it may not be fanciful to imagine more and more people following him, developing theories about abnormalities and testing them through sequencing. Such attempts will often fail, and in some cases lead to frustration and heartache. But some may make significant contributions to our understanding of the function of various human genes.

"I'm really being given an opportunity at this time in the history of genetics."
— Hugh Rienhoff

Rienhoff recognizes that he has benefited from his training and connections. But he told me part of his mission is to empower others. "I think probably the most important thing that people could take away from this is that the process is not mysterious," he says.

His enthusiasm is not universal. In the course of my reporting, Rienhoff gave his daughter's doctors permission to speak to me, and not all of them agreed that he was doing the right thing. Dietz says he worries that Rienhoff's example may lead some parents down the wrong path, searching for answers in the genes and diverting resources from the important goal of making sure their children are receiving proper care.

Rienhoff has heard these criticisms, and understands the discomfort. "There is a certain sense that all of this will unravel, meaning all of this will become driven by the people," he says. In deference to Dietz he has removed from his website a folder called 'How to sequence DNA' that he had never filled. "The purpose of the website is not about teaching people how to sequence DNA, at least not now," he says. But he still believes that patients and patient advocates can usher in what he calls a golden age in genetic research. It won't be for everyone. Rienhoff's search has been slow and methodical, and as yet inconclusive. Still, it has been fulfilling. "I'm really being given an opportunity, if you will, with this site and at this time in the history of genetics."

And the journey continues. Despite his daughter's DNA revealing not quite what he had expected, Rienhoff is still hopeful about his myostatin hypothesis, and it has led him to the most difficult decision he's ever made about his daughter. In May, based on the hypothesis that errantly activated signals might account for her inability to build muscles, Rienhoff, Hane and their daughter's cardiologist decided to put her on losartan, a drug for treating high blood pressure. Recent evidence suggests that it reduces the activity of secondary messengers triggered by TGF-β receptors[4], and that marfanoid mice are helped by the drug.

Rienhoff knows it is a controversial move, but there are two quite powerful factors on his pro list. First, if his daughter does have some aberrant form of Loeys–Dietz or Marfan's, the drug could forestall the vascular disease associated with the condition. Although he may eventually sequence her fibrillin and other genes for these disorders, the most definitive answers will come from regular cardiograms. The side effects of losartan are minor and reversible, he says, but vascular disease isn't. Second, her muscles might get a little better. "I tortured myself over that one," he says. "I took a Hippocratic oath — but I also took a parental oath — to do no harm."

Meanwhile he devours any literature on TGF-β signalling he can find. He has begun looking for scientists with whom he might collaborate on related projects, such as finding other patients with similar symptoms who might have mutations in the genes he's been looking at, or creating knock-out mice. He's waiting for more people to start using his website. He keeps a close eye on his daughter's progress, where he sees grounds for hope. "Her more proximal muscles seem to be growing," he says. "She can walk upstairs with a little assistance."

But he's cautious not to overinterpret. "I can't ascribe it to anything. I just keep my fingers crossed that she doesn't have a vascular disease. It's a quiet vigil." ■

Brendan Maher is a features editor at *Nature*.

1. Loeys, B. L. *et al. Nature Genet.* **37,** 275–281 (2005).
2. Dietz, H. C., Loeys, B., Carta, L. & Ramirez, F. *Am. J. Med. Genet.* **139C,** 4–9 (2005).
3. Schuelke, M. *et al. N. Engl. J. Med.* **350,** 2682–2688 (2004).
4. Habashi, J. P. *et al. Science* **312,** 117–121 (2007).

See Editorial, page 755.

NATURE|Vol 449|18 October 2007

SWERVE/ALAMY

WHAT'S IN THE RISING TIDE?

The nitrogen cycle rarely features in the grim litany of things at risk from global warming. **Nick Lane** reports on research that might change this — with grave consequences for ocean chemistry.

Colossal bridges bestride the waters of Narragansett Bay. Towns, interstate highways and suburbs sprawl along its shores and its waves are studded by thousands of pleasure boats. Yet the estuary retains a quiet beauty that befits the pious settlers who, 350 years ago, named the islands within the bay Prudence, Hope and Patience.

Extending its briny fingers through much of Rhode Island, Narragansett Bay has long juxtaposed man and nature, pollution and purity. Industrialization in the nineteenth century led to the rapid growth of cities such as Providence, a port at the bay's north end. The area now supports about a million people, all flushing their waste into the pastoral watershed. Reactive nitrogen compounds from treated sewage, industrial waste and fertilizers have poured in for decades, but remained at a relatively constant level for the past 25 years. Nevertheless, set against this steady background, a silent microbial and biochemical transformation has occurred in the bay that could have devastating ecological effects. The cause is pollution, but of an indirect sort — the changes seem to be down to global warming.

"Last year 17,000 tonnes of nitrate went unaccounted for in Narragansett Bay."

For the past three decades, the bottom sediments of the bay have mopped up much of the reactive nitrogen that humans have dumped into it. Although the fraction sequestered had been falling, this valuable natural sink has been protecting the bay and the coastal oceans from the effects of nitrate runoff. But last year the sediments abruptly stopped performing this service. Worse than that, they went into reverse. In a single summer, the bay switched from being a net sink to a net source of nitrates[1].

Robinson "Wally" Fulweiler, an oceanographer at Louisiana State University in Baton Rouge, and her colleagues at the University of Rhode Island in Kingston, were among the first to notice the turning environmental tide. If Narragansett is typical of other bays, they argue, it could be the harbinger of a new threat. Shifting the effect of anthropogenic nitrogen loading beyond the immediate coastal zone could destabilize ocean ecosystems by acidifying the waters, exacerbating harmful algal blooms, killing fish and shellfish, or perhaps even powering a vicious new cycle of global warming. The studies are currently hard to interpret and some say the system is poised to rebalance itself. But if they are wrong, global warming may do more to the oceans than make them rise.

According to Fulweiler, the root of the problem is a disconnect between life in the water column and that in the bottom sediments — the pelagic and benthic ecosystems, respectively. Under normal conditions, phytoplankton in the water column generate new organic matter through photosynthesis, some of which filters down to the bottom sediments — the benthos — where it provides food for bacteria. These benthic bacteria detoxify the water, removing excess nitrates, phosphates, and other pollutants while adding various micronutrients back to the water column. Healthy coastal ecosystems tend to have a good interchange — a tight coupling — between the pelagic and the benthic zones.

A silent switch

In Narragansett, the pelagic ecosystem is failing. Primary productivity, as measured by chlorophyll concentration, has fallen by 40% during the past 30 years, reflecting a dwindling of the spring bloom of phytoplankton.

Unlike harmful algal blooms — in which excess nutrients such as nitrates provoke a strangling growth of weed-like algae, ultimately leaving the water full of dead, rotting matter — normal seasonal blooms of phytoplankton are necessary to maintain the health of the estuary.

The decline, Fulweiler says, could have been caused by warmer winters, either because they provide thicker cloud cover or because they allow more grazing zooplankton to flourish. Either way, rising temperatures are hitting primary productivity.

Falling productivity means a decline in the quantity and quality of organic matter reaching benthic bacteria, notably the denitrifiers. Denitrifying bacteria convert nitrates back to inert nitrogen gas. Just as organic molecules from food are reacted with oxygen to generate energy in animals, these denitrifying bacteria glean energy when organic remains react with nitrate.

The benthic denitrifiers are choosy eaters. They live on 'labile organic carbon', essentially, fresh food. A decline in primary productivity was likely to hit them hard. What came as a shock was the switch to a completely new population — the nitrogen-fixing bacteria. These organisms take nitrogen dissolved in water and convert it into ammonia, in a process known as nitrogen fixation. This new organic nitrogen is ultimately converted into nitrates by a third group of bacteria, the nitrifiers. So the sediments not only stop mopping up excess nitrates, they start adding more to the pot (see 'The changing cycle').

What flipped the switch is unknown. The nitrifying bacteria at the end of this chain have been shown to have unpredictable population patterns[2]. If that is the case here, then the change may not be significant. But a more troubling explanation relates to the composition of non-labile organic matter. According to Bess Ward, a biogeochemist at Princeton University in New Jersey, organic matter can become so depleted of nitrogen that it no longer provides enough sustenance for denitrifying bacteria. That stops their growth. Because nitrogen fixers don't face this constraint, they thrive.

If Ward is right, then the switch could be both persistent and widespread. The overall equation is not trivial. Under normal circumstances, the sediments in Narragansett Bay decontaminate around one quarter of the reactive nitrogen compounds running off from farmland and sewage — something in the order of 1,000 tonnes of nitrogen, or 5,000 tonnes of nitrate, every year — making the failure of denitrification significant in its own right. Fulweiler notes that the excess nitrogen is not accumulating as dissolved nitrates in the bay, nor is it stimulating algal blooms. Presumably, she says, at least some is simply being flushed out to sea.

S. NIXON

"Presumably, some of the excess nitrogen is being flushed out to sea."
— Robinson "Wally" Fulweiler

An unbalanced equation

In addition to this substantial flux, in the three summer months of 2006, the rate of nitrogen fixation by the thriving nitrogen fixers was estimated to be around 1.5 times greater than the total input from rivers, sewage and atmospheric pollution combined — nearly 3,000 tonnes of nitrogen converting into 12,000 tonnes of nitrate. Even allowing for more typical conditions throughout the rest of the year, this still represents 20–60% more nitrogen input annually. Last year, in Narragansett Bay, some 17,000 tonnes of nitrate went unaccounted for. So, what's happening to it all?

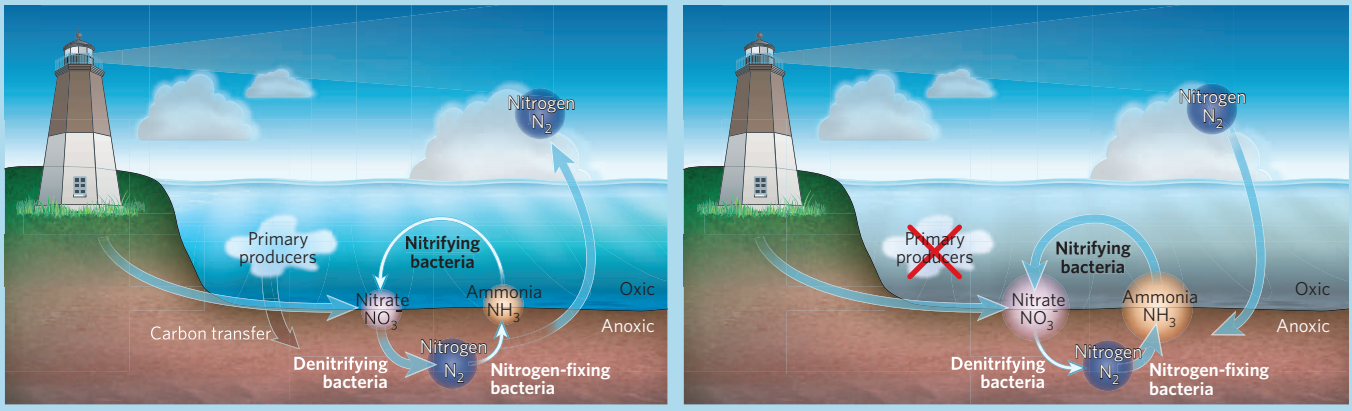
A switch from denitrification to nitrogen fixation is utterly unexpected, which in itself shows just how much remains to be learned about the nitrogen cycle. Fulweiler notes that nitrogen-fixing bacteria in estuarine sediments have not been seen as important players in the nitrogen cycle before. But balancing the nitrogen cycle has been a bit of an embarrassment for years. It should be easy enough — just sum up the total nitrate inputs, from natural and anthropogenic sources, and subtract the flux back to the atmosphere as nitrogen gas. Instead, it's a giant mismatch. The calculated rate of global denitrification is twice the known inputs from fixation and anthropogenic sources. As Lou Codispoti at the University of Maryland's Horn Point Laboratory (HPL) just outside Cambridge says, either the cycle doesn't balance at all — that is, today's oceans are in some sort of a 'transient' state — or scientists have overlooked a whole lot of nitrogen fixation somewhere.

Most suspect the latter. And set against this background, Narragansett Bay might just represent a small fraction of that missing nitrogen. On the other hand, if global warming really does alter the nitrogen cycle, it could throw the global equation off balance.

Fulweiler is the first to admit that a correlation is not proof of causality. Nevertheless, she's concerned with what might be a trend. There are hints from elsewhere that nitrogen fixation is picking up. A similar uncoupling of the pelagic and benthic ecosystems was noted in the arctic last year. Jackie Grebmeier at the University of Tennessee in Knoxville, and her collaborators reported[3] a 75% drop in benthic oxygen consumption (a surrogate for carbon supply) between 1988 and 2004. More recently, oceanographer Judy O'Neil at the

THE CHANGING CYCLE

When primary producers in Narragansett Bay started to dwindle, the benthic balance of bacteria shifted in favour of nitrogen-fixing bacteria.



HPL measured nitrogen fixation in tributaries to Chesapeake Bay, a few hundred miles south of Narragansett. Although she's yet to quantify their total contribution, her data incriminate cyanobacterial blooms, rather than benthic nitrogen-fixation for flipping the switch. The reason is that cyanobacterial blooms are notoriously full of toxins, which makes them unpalatable to grazing zooplankton, and the animals that eat them in turn. As a result, more carbon is flushed out in a dissolved state rather than passing down to the bottom sediments via faecal pellets.

The trend appears new according to Ward, who, working with Todd Kana and his colleagues at the HPL, looked for evidence of nitrogen fixation in the Chesapeake 5 years ago; they found no activity. Later, working with Jon Zehr and others at the University of California, Santa Cruz, Ward found copies of the genes required for fixation in the sediments there[4]. What's more, they were associated with the most unusual suspects, such as proteobacteria, which have never before been associated with nitrogen fixation. Nitrogen fixation requires a suite of 10 to 20 genes; if lying fallow, these are costly to maintain and so ought to be lost. Ward says that if they're there, they're being used.

Some like it hot

In the Chesapeake, too, the rise in nitrogen fixation probably relates to global warming, albeit for different reasons. Here there is no evidence of failing spring blooms, but rather a gradual takeover by cyanobacteria, which apparently "like it hot". Don Canfield, a geochemist at the University of Southern Denmark in Odense, says he thinks the trend is global. If he's right, nitrate export could rise substantially in a matter of a few years. Of course verifying the export of nitrogen hinges on balancing that stubborn nitrogen cycle equation. Paul Falkowski, a biogeochemist at Rutgers University's Institute of Marine and Coastal Sciences in New Brunswick, New Jersey, is sceptical about nitrogen export from Narragansett Bay. If ammonia and nitrates are being produced in such large quantities, he asks, then why are we seeing a steady decline in primary productivity, rather than an increase in blooms? And if the nitrates are all being flushed out into coastal waters, then the more buoyant, fresher water from the bay (with its river inputs) ought to float in the sunny surface waters above the denser saline, and we should see blooms there instead. Satellite images, he says, don't show dramatic blooms.

Fulweiler has plausible answers, which need to be tested: in the bay itself, nitrates may be swallowed up by other bacteria competing for resources with phytoplankton in the water column. If so, then the export of nitrates to the oceans would be more limited, and the problem would revert back to one of local ecology. On the other hand if more nitrates really are exported out to sea, Fulweiler argues, they wouldn't necessarily stimulate large algal blooms. The salinity gap between Narragansett Bay and the ocean is quite small, she says, so the waters emerging from the bay might not float.

In the absence of dramatic blooms, the water might instead be getting more acidic, according to Scott Doney and his collaborators at Woods Hole Oceanographic Institute in Massachusetts. Both nitrates and sulphates acidify ocean waters. The impact on the open ocean is limited, compared with the acidification caused by rising carbon dioxide levels, but Doney's marine modelling suggests that the effects of nitrates in coastal waters may be serious. The organisms most likely to be affected by ocean acidification are those with shells or skeletons made from calcium carbonate, including many algae that would otherwise bloom[5].

"Either the oceans are in some sort of transient state or scientists have overlooked a lot of nitrogen fixation."
— Lou Codispoti

No laughing matter

Another possibility, if the waters emerging from Narragansett sink, is a bloom of denitrifying bacteria lower in the water column. Denitrification in coastal waters tends to be dominated by the classic bacterial pathway; and a major by-product of this is nitrous oxide — laughing gas. Nitrous oxide is 200–300 times more potent as a greenhouse gas than carbon dioxide.

In 2000, Wajih Naqvi and his colleagues at the National Institute of Oceanography in Goa, India, reported an alarming accumulation of nitrous oxide in the Arabian Sea, along the western Indian continental shelf, following a monsoon washout of nitrate fertilizers[6]. The calculated emissions from the region in just 6 months accounted for as much as 5% of annual ocean emissions from a region that makes up only 0.05% of the world's oceans. Although there is much to learn about the interplay of factors controlling oceanic nitrous oxide emissions, Naqvi says that emissions are generally greatest in oxygen-minimum zones. Because these are spreading as a result of global warming (oxygen is less soluble in warmer waters) and eutrophic conditions (an increase in chemical nutrients) nitrous oxide emissions will in all probability rise.

Most observers anticipate that restrictions in the use of nitrates, along with better facilities for treating sewage and industrial waste, will reduce oceanic nitrogen contamination. But a widespread switch to nitrogen fixation, as a result of global warming, could raise nitrate levels and nitrous oxide emissions despite human intervention, driving a vicious cycle. Balancing the nitrogen cycle as the world warms suddenly looks more critical and more uncertain than ever. ■

Nick Lane is a science writer and honorary reader at University College London.

Judy O'Neil samples a thick bloom of assorted algae (inset) from the Sassafras River, a tributary of Chesapeake Bay in Maryland.

B. DENNISON

1. Fulweiler, R. W., Nixon, S. W., Buckley, B. A. & Granger, S. L. *Nature* **448,** 180–182 (2007).
2. Graham, D. W. *et al. ISME Journal* **1,** 385–393 (2007).
3. Grebmeier, J. M. *et al. Science* **311,** 1461–1464 (2006).
4. Jenkins, B. D., Steward, G. F., Short, S. M., Ward, B. B. & Zehr, J. P. *Appl. Environ. Microbiol.* **70,** 1767–1776 (2004).
5. Doney, S. C. *et al. Proc. Natl Acad. Sci. USA* **104,** 14580–14585 (2007).
6. Naqvi, S. W. *et al. Nature* **408,** 346–349 (2000).

Geo-engineering might cause, not cure, problems

SIR — James E. Lovelock and Chris G. Rapley, in their Correspondence 'Ocean pipes could help the Earth to cure itself' (*Nature* **449,** 403; 2007) propose a variant on some well-publicized schemes to remove carbon dioxide from the atmosphere, by fertilizing the surface waters of the ocean (see also *Nature* doi:10.1038/news070924-8; 2007). All such schemes suffer from a major problem, because simply enhancing the growth of phytoplankton is not enough. It is the sinking flux of particulate organic carbon into the deep ocean — and ideally into the sediments (usually a small fraction of the total primary production) — that must be enhanced for sequestration to be effective.

Several recent open-ocean experiments have attempted to quantify the level of growth enhancement and sequestration caused by purposeful fertilization: for example, by iron, an essential micro-nutrient. Despite successfully increasing plant biomass, these have not demonstrated sequestration of carbon into the deep ocean (below 1,000 metres), which is essential if it is to be isolated from the atmosphere for centuries or longer. The sinking particles carrying this carbon are degraded rapidly by respiration, and mostly remineralized within the upper ocean. It is likely that almost all the CO_2 taken up is released back to the atmosphere within a year. Also, this scheme would bring water with high natural p_{CO_2} levels (associated with the nutrients) back to the surface, potentially causing exhalation of CO_2.

We support efforts to find ways of sequestering carbon, but the likely consequences of geo-engineering schemes should be thoroughly researched before they are promoted as solutions. We do not consider ocean fertilization to be a promising approach, and on a large scale it would constitute major interference with an ecosystem which is still poorly understood. Fertilization is likely to alter the phytoplankton community composition and succession, and thus the structure of the oceans' food webs. It might damage these remote and possibly fragile ecosystems, trigger unexpected feedbacks and even reduce their ability to sequester carbon. We cannot, therefore, support this approach, until it can be shown that there would be demonstrable benefits which would outweigh the potential impacts.

John Shepherd, Debora Iglesias-Rodriguez, Andrew Yool
National Oceanography Centre, Southampton, SO14 3ZH, UK

See Nature Reports Climate Feedback http://blogs.nature.com/climatefeedback/2007/09/lovelock_and_rapley_propose_cu_1.html and the *Nature* newsblog at http://blogs.nature.com/news/blog/2007/09/mixing_the_oceans_proposed_to.html for further comments on Lovelock and Rapley's Correspondence; you are welcome to add your own views — Editor, *Nature*.

Heavy workload may have led to mistakes in review

SIR — Michael M. Crow, president of Arizona State University, is renowned for striving to make his institution the biggest in the United States while raising its relatively low academic standing. His impartiality may be open to question as a reviewer of my book *Science for Sale: The Perils, Rewards, and Delusions of Campus Capitalism* (*Nature* **449,** 405; 2007), which is sceptical of such quests. Its epilogue, 'A parable for our time', caricatures the headlong pursuit of academic greatness.

Not open to question, however, are Crow's misrepresentations. Nowhere do I assert "that the academic scientist and the university are best motivated by curiosity alone". I do report an interview I conducted in which a scientist makes that assertion, followed by my rejoinder challenging the concept. And, contrary to Crow's assessment, I do not ignore or disparage the long history of practical research in universities: I go into considerable detail on this subject.

I prefer to believe that hasty reading by a heavily burdened university president accounts for these errors and omissions.

Daniel S. Greenberg
3736 Kanawha St. Nw.,
Washington DC 20015-1874, USA

Funding basic research brings unexpected benefits

SIR — The United Kingdom's research base has seen unprecedented increases in public investment in recent years, mostly predicated on the long-term benefits to society expected to arise from that investment. It is the research councils' responsibility, as the major public funders of UK research, to provide compelling evidence that these expectations are being met. Your Editorial 'Innovation versus science?' (*Nature* **448,** 839; 2007) concludes that efforts to document this herald a shift away from our support for basic research. As a research council chief executive, leading our efforts to increase our economic impact, I can say that is not the case.

The UK Research Councils have just published a report, *Excellence with Impact* (www.rcuk.ac.uk/innovation/impact) that looks across research councils' investments. Each of 18 case studies shows actual and/or potential impact, ranging from biotech spin-outs and skilled engineers to climate-change policy. Probably the most reassuring finding was the extent to which some demonstrated multiple types of impact. Furthermore, many of the impacts were not necessarily part of the original rationale for the specific investment, suggesting that serendipity and opportunism are important factors for the research councils. Investment in DNA technologies, for example, did not anticipate the forensic power of DNA fingerprinting, and polymer research was not funded with the anticipation that it would create a new market in flexible displays.

These results demonstrate the wisdom of the research councils' commitment to funding excellent basic research. Rather than weaken that commitment, our approach is to embed economic-impact considerations in our organizations, thus shifting the central focus of the research councils to excellent research with high economic impact. So it is about what basic research we should fund, rather than if we should fund it.

Philip Esler
Arts and Humanities Research Council, Whitefriars, Lewins Mead, Bristol BS1 2AE, UK

One-vesicle hypothesis has been extensively discussed

SIR — Your News story 'Long-held theory is in danger of losing its nerve'[1] described a published criticism of work we published 25 to 26 years ago and our reply (references are in ref. 1). In it, you quote unnamed experts who maintain that much of the published work that might be consistent with the one-vesicle hypothesis addressed in the News story has problems.

Recognizing that this News story was not a scientific article, we think it is important to clarify that our work, and that hypothesis, have been discussed extensively in the literature and in public forums[2], including the exchange of data and analytical tools with another laboratory (see ref. 3). Data consistent with it have also been published quite recently[4]. We all now recognize that neurotransmitter chemical release at synapses is heterogeneous and may function differently in different biological systems.

Henri Korn*, Donald Faber†, Antoine Triller‡, Alain Mallet§
*INSERM, Institut Pasteur, Paris, France
†Department of Neuroscience, Albert Einstein College of Medicine, Bronx, New York, USA
‡Biologie cellulaire de la Synapse, Inserm U497, Ecole Normale Supérieure, 75005 Paris, France
§Département de Biostatistiques et Biomathématiques, Université Pierre et Marie Curie–Paris 6, France

1. *Nature* **449,** 124–125 (2007).
2. *Central Synapses: Quantal Mechanisms and Plasticity* (eds. Faber, D. S. *et al.*) HFSP, Strasbourg (1998).
3. Redman, S. *Physiol. Rev.* **70,** 165–198 (1990).
4. Silver, R. A. *et al. Science* **302,** 1981–1984 (2003).

Geo-engineering might cause, not cure, problems

SIR — James E. Lovelock and Chris G. Rapley, in their Correspondence 'Ocean pipes could help the Earth to cure itself' (*Nature* **449,** 403; 2007) propose a variant on some well-publicized schemes to remove carbon dioxide from the atmosphere, by fertilizing the surface waters of the ocean (see also *Nature* doi:10.1038/news070924-8; 2007). All such schemes suffer from a major problem, because simply enhancing the growth of phytoplankton is not enough. It is the sinking flux of particulate organic carbon into the deep ocean — and ideally into the sediments (usually a small fraction of the total primary production) — that must be enhanced for sequestration to be effective.

Several recent open-ocean experiments have attempted to quantify the level of growth enhancement and sequestration caused by purposeful fertilization: for example, by iron, an essential micro-nutrient. Despite successfully increasing plant biomass, these have not demonstrated sequestration of carbon into the deep ocean (below 1,000 metres), which is essential if it is to be isolated from the atmosphere for centuries or longer. The sinking particles carrying this carbon are degraded rapidly by respiration, and mostly remineralized within the upper ocean. It is likely that almost all the CO_2 taken up is released back to the atmosphere within a year. Also, this scheme would bring water with high natural p_{CO_2} levels (associated with the nutrients) back to the surface, potentially causing exhalation of CO_2.

We support efforts to find ways of sequestering carbon, but the likely consequences of geo-engineering schemes should be thoroughly researched before they are promoted as solutions. We do not consider ocean fertilization to be a promising approach, and on a large scale it would constitute major interference with an ecosystem which is still poorly understood. Fertilization is likely to alter the phytoplankton community composition and succession, and thus the structure of the oceans' food webs. It might damage these remote and possibly fragile ecosystems, trigger unexpected feedbacks and even reduce their ability to sequester carbon. We cannot, therefore, support this approach, until it can be shown that there would be demonstrable benefits which would outweigh the potential impacts.

John Shepherd, Debora Iglesias-Rodriguez, Andrew Yool
National Oceanography Centre, Southampton, SO14 3ZH, UK

See Nature Reports Climate Feedback http://blogs.nature.com/climatefeedback/2007/09/lovelock_and_rapley_propose_cu_1.html and the *Nature* newsblog at http://blogs.nature.com/news/blog/2007/09/mixing_the_oceans_proposed_to.html for further comments on Lovelock and Rapley's Correspondence; you are welcome to add your own views — Editor, *Nature*.

Heavy workload may have led to mistakes in review

SIR — Michael M. Crow, president of Arizona State University, is renowned for striving to make his institution the biggest in the United States while raising its relatively low academic standing. His impartiality may be open to question as a reviewer of my book *Science for Sale: The Perils, Rewards, and Delusions of Campus Capitalism* (*Nature* **449,** 405; 2007), which is sceptical of such quests. Its epilogue, 'A parable for our time', caricatures the headlong pursuit of academic greatness.

Not open to question, however, are Crow's misrepresentations. Nowhere do I assert "that the academic scientist and the university are best motivated by curiosity alone". I do report an interview I conducted in which a scientist makes that assertion, followed by my rejoinder challenging the concept. And, contrary to Crow's assessment, I do not ignore or disparage the long history of practical research in universities: I go into considerable detail on this subject.

I prefer to believe that hasty reading by a heavily burdened university president accounts for these errors and omissions.

Daniel S. Greenberg
3736 Kanawha St. Nw.,
Washington DC 20015-1874, USA

Funding basic research brings unexpected benefits

SIR — The United Kingdom's research base has seen unprecedented increases in public investment in recent years, mostly predicated on the long-term benefits to society expected to arise from that investment. It is the research councils' responsibility, as the major public funders of UK research, to provide compelling evidence that these expectations are being met. Your Editorial 'Innovation versus science?' (*Nature* **448,** 839; 2007) concludes that efforts to document this herald a shift away from our support for basic research. As a research council chief executive, leading our efforts to increase our economic impact, I can say that is not the case.

The UK Research Councils have just published a report, *Excellence with Impact* (www.rcuk.ac.uk/innovation/impact) that looks across research councils' investments. Each of 18 case studies shows actual and/or potential impact, ranging from biotech spin-outs and skilled engineers to climate-change policy. Probably the most reassuring finding was the extent to which some demonstrated multiple types of impact. Furthermore, many of the impacts were not necessarily part of the original rationale for the specific investment, suggesting that serendipity and opportunism are important factors for the research councils. Investment in DNA technologies, for example, did not anticipate the forensic power of DNA fingerprinting, and polymer research was not funded with the anticipation that it would create a new market in flexible displays.

These results demonstrate the wisdom of the research councils' commitment to funding excellent basic research. Rather than weaken that commitment, our approach is to embed economic-impact considerations in our organizations, thus shifting the central focus of the research councils to excellent research with high economic impact. So it is about what basic research we should fund, rather than if we should fund it.

Philip Esler
Arts and Humanities Research Council, Whitefriars, Lewins Mead, Bristol BS1 2AE, UK

One-vesicle hypothesis has been extensively discussed

SIR — Your News story 'Long-held theory is in danger of losing its nerve'[1] described a published criticism of work we published 25 to 26 years ago and our reply (references are in ref. 1). In it, you quote unnamed experts who maintain that much of the published work that might be consistent with the one-vesicle hypothesis addressed in the News story has problems.

Recognizing that this News story was not a scientific article, we think it is important to clarify that our work, and that hypothesis, have been discussed extensively in the literature and in public forums[2], including the exchange of data and analytical tools with another laboratory (see ref. 3). Data consistent with it have also been published quite recently[4]. We all now recognize that neurotransmitter chemical release at synapses is heterogeneous and may function differently in different biological systems.

Henri Korn*, Donald Faber†, Antoine Triller‡, Alain Mallet§
*INSERM, Institut Pasteur, Paris, France
†Department of Neuroscience, Albert Einstein College of Medicine, Bronx, New York, USA
‡Biologie cellulaire de la Synapse, Inserm U497, Ecole Normale Supérieure, 75005 Paris, France
§Département de Biostatistiques et Biomathématiques, Université Pierre et Marie Curie–Paris 6, France

1. *Nature* **449,** 124–125 (2007).
2. *Central Synapses: Quantal Mechanisms and Plasticity* (eds. Faber, D. S. *et al.*) HFSP, Strasbourg (1998).
3. Redman, S. *Physiol. Rev.* **70,** 165–198 (1990).
4. Silver, R. A. *et al. Science* **302,** 1981–1984 (2003).

COMMENTARY

Common sense for our genomes

A personal DNA sequence is not yet practically useful. But it could be, argues **Steven E. Brenner**, if we had the right resources available to interpret genomes.

Revelation of the complete DNA sequences of James Watson and J. Craig Venter elicited headlines in recent months, but most press reports struggled to offer meaningful interpretations. The most noted observation was that Venter has a particular gene variant predisposing him to cardiac disease, although his family history was enough to let him know about this general risk. If the genome is so revealing, why was so little revealed?

It is telling that Venter said he learned about the cardiac disease gene in a newspaper report. Put simply, even we in the scientific community can't easily come to grips with what we know. The effects of gene variations are scattered in hundreds of databases, across hundreds of interpretative reports in clinical laboratories, and among millions of manuscripts and patent applications. And although some papers discuss the precise effects of a single DNA base change, many analyses offer simple rules of thumb rather than specific guidance.

Moreover, even as we celebrate the advent of personal genome sequencing, we should maintain realistic expectations. Given that most common drug prescriptions don't even consider a patient's weight, it is unclear how many future therapies will depend on the minutiae of our genomic make-up. Indeed, it remains to be seen whether we will typically learn anything more important from our genomes than the need to use sunscreen, eat better and exercise more. However, I believe that if we don't seize the initiative and develop the necessary resources to interpret our genomes, the Venter and Watson genomes will be seen as missed opportunities.

Even the scientific paper reporting Venter's genome revealed less than it might[1]. The gene variants described in the initial analysis, intended to engage a wider audience, could have been selected to elicit guffaws, touching on associations with alcoholism, obesity, novelty-seeking and antisocial behaviour. However, these are all statistical likelihoods, and their relevances are hard to decipher.

Yet, after learning of the genetic variations that render him susceptible to cardiac disease, Craig Venter reportedly assumed a new level of personal responsibility by altering his diet and taking a cholesterol-lowering statin. So personal genomes may offer a way to translate genomic knowledge into better preventive medicine.

Even now, further analyses of the Venter genome[2] could reveal more useful gene variants.

R. WOODWARD/GETTY

For example, cytochrome P450 isozymes determine how rapidly individuals metabolize various drugs, and the US Food and Drug Administraton has approved a microarray test for genotyping these enzymes. Venter's cytochrome P450 gene variants were not reported, but these variations can inform drug dosages.

We are still waiting to learn if the analysis of Watson's genome will reveal more or less than Venter's. Watson's sequence is available online[3] and a small number of gene variants have been automatically annotated using the Online Mendelian Inheritance in Man (OMIM) database. OMIM has 18,000 entries summarizing the literature related to human genes and genetic disorders (see table overleaf). But because such mutations and their effects are described textually, only 133 of the 18,000 could be linked directly to a unique single-nucleotide substitution[4].

"It remains to be seen whether we will learn anything more important from our genomes than the need to use sunscreen, eat better and exercise more."

Visionary geneticists have long contemplated building a resource to consolidate our understanding of genome variation. However, academic squabbles and misunderstandings caused the most comprehensive effort — involving hundreds of scientists backed with millions of dollars — to founder[5]. Perhaps they were premature? Until recently, it was rarely productive to look beyond a single gene known to be of research or clinical interest. Today, the situation has changed radically. With the prospect of inexpensive personal genome sequences, there is profound impetus for integrating our knowledge of genetic variation and its effect on a genomic scale.

Covering the bases

Many of the foundations for describing human genome variation and integrating this knowledge are already in place. The Human Genome Variation Society has defined a standard nomenclature for precisely describing small variants, which makes it possible, for example, to consistently ascertain whether two polymorphisms are the same or different. Central publicly funded databases have repositories of genetic variation information and offer reference genes and genomes on which the variation can be mapped. Among these, dbGaP is an example of a database of genotype–phenotype relationships generated largely from genome-wide association studies. There are also more than 600 locus-specific databases that focus on narrow areas of the genome. But merging these databases with dbGaP, and other data sources, would be a complex task.

I propose establishing a Genome Commons, a public knowledgebase of human genetic variation and its effect, culled from databases, diagnostic laboratories, and the scientific literature. Ultimately, such a repository of our common human inheritance would be a vast resource for research, medicine and understanding ourselves.

There are many ways in which the Genome Commons could be constructed, but I offer some general guiding principles. It would certainly build on the curation of hundreds of small locus-specific and other databases today. This is an often used and successful model, employed for example at GeneTests,

SOME EXISTING SOURCES FOR INTERPRETING HUMAN GENOMES

Name	Website	Brief description	Restrictions on use
dbSNP	www.ncbi.nlm.nih.gov/SNP/	Repository for short nucleotide polymorphisms	None
OMIM, Online Mendelian Inheritance in Man	www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM	Catalogue of 18,000 essays on human genes and genetic disorders	Licence for commercial use or redistribution
dbGaP	www.ncbi.nlm.nih.gov/sites/entrez?db=gap	Mainly a database from genome-wide association studies	None on open data, some on personal
SNPedia	www.snpedia.com	Wikipedia-style site for single-nucleotide polymorphisms	None
HGMD, Human Gene Mutation Database	www.hgmd.cf.ac.uk/ac/index.php	Catalogue of gene mutations responsible for human inherited disease	Fee-based for full access; no redistribution
GeneTests	www.genetests.org/	Summarizes more than 1,000 diagnostic genetic tests	None with proper attribution
PharmGKB	www.pharmgkb.org/	Pharmacogenetics and pharmacogenomics knowledgebase	Some privacy restrictions
Locus Specific Mutation Databases	www.hgvs.org/dblist/glsdb.html	Lists over 600 locus specific databases	Some copyright restrictions
SIFT	http://blocks.fhcrc.org/sift/SIFT.html	Software predicting sequence effects on protein function	None with proper attribution
SNPs3D	www.snps3d.org	Website that predicts phenotypic impact of SNPs	Software not downloadable

A more comprehensive list is compiled by Rania Horaitis and available on the Human Genome Variation Society (HGVS) website at http://www.hgvs.org/dblist/dblist.html

a reference database of thousands of gene and disease tests for diagnostic use. The editors of GeneTests benefit from contributions by hundreds of experts who volunteer their knowledge. Similarly, quality controls in the Genome Commons would be provided by experts overseeing entries in their domain of expertise, typically a set of genes or diseases. In addition to their own contributions, they would collate and review entries that could be submitted by anyone with access to academic journals and appropriate training.

Share and share alike

To work on a genomic scale, the Genome Commons would need to be carefully structured, incorporating statistical details about data quality and the strength of associations for researchers, as well as clinical references for eventual use by medical practitioners. It is essential that the Genome Commons be open for remixing, augmentation and redistribution of content. It is only in this way that researchers can fully share their knowledge and allow others to build on it.

An individual genome will typically have millions of differences when compared with a reference genome; most differences are of little consequence, but some single mutations can be fatal. The Genome Commons itself need not contain any individual's information and thus raises few ethical or privacy concerns. However, both for research purposes and for clinical interpretation, we will need a navigation tool to relate each individual's variations to the knowledge compiled in the Genome Commons.

But sequenced genomes do not come indexed for easy analysis and our knowledge is so multilayered, that it presents a technical challenge. At one extreme, for sickle-cell anaemia, we understand the molecular mechanism by which mutation leads to disease. In many more instances, however, there is a single-gene association, without any mechanistic understanding. In general, we are happy to find any significant association of phenotype with a genetic marker. Most variations have never been phenotypically characterized — Venter's genome had more than a million variants never seen before — and analysing these will require predictive approaches. Moreover, variations appear on different scales in the genome, ranging from small substitutions, insertions and deletions, to large-scale chromosomal restructuring.

Initially, I imagine that a Genome Commons navigator would amalgamate observed variation, and propose phenotypic interpretations. This first step would allow researchers to assess the challenge and promise of these data, and to design further research and analysis methods. Later versions of navigators will incorporate the best methods from many research groups. But to truly interpret a genome, we face the more daunting challenge of sifting through the millions of variations and ranking them so that we are not deluged with genomic marginalia. The navigator would eventually present a status report focusing on genetic differences of greatest medical or personal importance.

Private enterprise would play a vital part by providing an interface between the Genome Commons and the wider community. Researchers would access the Genome Commons directly, but companies would mediate its delivery to patients and physicians. Just as clinical laboratories are used by physicians to perform diagnostic testing today, I would expect clinical labs to perform large-scale genome sequencing in the future. I envisage these labs — and new companies such as 23andMe and Navigenics — using the Genome Commons navigator as a reference tool for producing diagnostic reports.

Much genomic variation information is not free, or is encumbered with intellectual-property protection. To be fully successful, companies must also contribute discoveries to the Genome Commons. As a central clearing house of intellectual property, the Genome Commons could reduce transaction costs. Companies could contribute information and accept a standard agreement for diagnostic use, making it easier for clinical laboratories to license large quantities of intellectual property with minimal overheads. In this way, more assays become accessible and affordable to patients.

The cost to create and maintain the Genome Commons will be considerable, even if many volunteers assist the effort. Extrapolating from the costs of other resources, such as OMIM, PharmGKB and GeneTests, the core knowledgebase may require millions of dollars in support each year. Most of this would be spent on salaries for curators and staff overseeing the informatics.

Ideally, the Genome Commons would be primarily funded as a government resource or by a major charity, although many companies will have strategic economic reasons to financially support an open resource. If a public Genome Commons fails to emerge, we may instead get a private resource with similar content, but whose licensing requirements stymie research and innovation. A single private resource would also lead to monopoly pricing for diagnostic information. After the huge investments made to ensure that a human genome sequence was public and free, additional outlays for the Genome Commons seem prudent so that genomes can be readily interpreted for medical practice and research.

The challenges of building a Genome Commons and navigator are not trivial, but this resource could affect us all personally. In a world where we all face limited time, resources and personal restraint, an open Genome Commons would eventually enable productive use of the wealth of information available, helping us to prioritize healthy activities and therapies to give us the most productive and enjoyable lifespans. ■

Steven E. Brenner is at the Department of Plant and Microbial Biology, 111 Koshland Hall, University of California, Berkeley, California 94720, USA.

1. Levy S, *et al. PLoS Biol.* **5,** e254 (2007).
2. www.jcvi.org/research/huref/
3. http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/
4. https://mice.cs.columbia.edu/getTechreport.php?techreportID=448&format=pdf
5. Maurer, S. M. *Res. Policy* **35,** 839–853 (2006).

Join the discussion at www.GenomeCommons.org

BOOKS & ARTS

A life worth writing about

Craig Venter's autobiography recounts the conflict and controversy that have contributed to his celebrity.

A Life Decoded: My Genome: My Life
by Craig Venter
Allen Lane/Viking: 2007. 400 pp.
£25/$25.95

Jan Witkowski

Reviewing Jim Watson's *The Double Helix*, Erwin Chargaff dismissed scientific autobiography as "a most awkward literary genre", arguing that most scientists lead monotonous and uneventful lives. This certainly does not apply to Craig Venter, whose autobiography is fittingly well-written, fast-paced and full of interesting data, gossip — and score-settling.

Little introduction is necessary for the man who is possibly the celebrity scientist of the modern era. Venter's name has rarely been out of the headlines for the past 12 years. Most recently, on the publication of his own genome sequence, his portrait occupied one-third of the front page of the *New York Times*'s science section. His fame peaked at the beginning of the millennium at the celebrations for the first release of the human-genome sequence. Inspired by Darwin's *Beagle* voyage, he then set sail in his yacht *Sorcerer II* to catalogue the oceans' bacteria and viruses.

C. VENTER

Craig Venter on his yacht, the venue for a gargantuan project to catalogue the oceans' bacteria and viruses.

Close encounter

A Life Decoded is in three parts, divided by two epiphanies. Venter grew up a bad boy in Millbrae near San Francisco. On his bike he raced planes taking off from the airport, and hung out with other bad boys. Many character traits formed early, notably the competitive — and combative — spirit: his track relay team set a US record and his swimming coach thought him to be potential Olympic material. His love of sailing began when he built a hydroplane to plans in *Popular Mechanics*. That Venter graduated was a close-run thing: his grades were poor and he led a school sit-in to protest against a teacher's dismissal. At 17, he set off to southern California for surfing, girls and drink.

In late 1960s America, the good times couldn't last for a healthy young man. The armed forces beckoned and Venter enlisted in the hospital corps. His talent for spinal taps kept his name off the lists of postings to Vietnam. But he was eventually shipped off to a hospital in Da Nang. Experiences of working with the wounded and dying left a lifelong mark on Venter, of which he writes with candour and passion.

Epiphany number one came after 5 months in Vietnam. He decided to commit suicide by swimming out to sea. Out of sight of land, a nudge from a shark made him race back. He served out his time in Vietnam and returned to California in late 1968, marrying his girlfriend and buying a motorbike en route.

Venter quickly made up for his misspent youth, enrolling first in a community college, then moving to the University of California, San Diego. He studied biochemistry and researched into the action of adrenaline. Publishing his work proved more satisfying than winning swimming or sailing competitions. After graduation, Venter continued with research on hormones and their receptors, completing his doctoral thesis in 1975. Following a productive period at the State University of New York, Buffalo, where his son was born, he divorced his wife and married Claire Fraser, now director of the Institute for Genome Sciences at the University of Maryland School of Medicine in Baltimore. In 1984, he and Fraser moved to the National Institute of Neurological Disorders and Stroke (NINDS) in Bethesda, Maryland, and continued to work on neurotransmitter receptors.

Epiphany number two came in 1987. Venter read a paper from Lee Hood's laboratory at the California Institute of Technology on the automation of DNA sequencing. After months of toil to clone and sequence the human brain β-adrenergic receptor, the idea that a machine could replace postdocs and technicians was thrilling. He phoned Mike Hunkapiller, who developed the automated sequencers at CalTech, and ordered a machine. In a prophetic turn of events, he paid for it with his rainy-day savings when his NINDS boss allegedly refused to fund it. Within a few months, Venter and his colleagues had sequenced two receptors.

It was about this time that I met Venter at a conference on new developments in complex human genetic disorders, where he displayed his characteristic impatience with those who play it safe. One participant remarked that what would most help her research would be the sequence of the 300-kilobase region she was studying. Venter said, in effect, why don't you just get on and do it then? His suggestion was regarded with incredulity.

Genome giant

The larger part of *A Life Decoded* chronicles Venter's controversial move into genomics and his part in sequencing the human genome. He covers a lot of ground, with facts, quotations and accounts of others' intentions. Some assertions lack references, so the reader must take them on trust. Venter pillories a long list of people whom he believes conspired and lied to obstruct him, but offers generous praise for the staff of his institutes who tackled obstacles not of their making.

Clashes began while Venter was at the National Institutes of Health (NIH) in the early nineties, over sequence patenting and strategies for cracking whole genomes. Eventually, Venter left and set up The Institute for Genomic Research in Rockville, Maryland. His team used whole-genome assembly (WGA) to

sequence the *Haemophilus influenzae* genome, which was eight times larger than the previous record. Instead of painstakingly cloning, mapping the clones and sequencing them, they blasted *H. influenzae* DNA into fragments, sequenced these, and used a computer to assemble the complete genome that graced the cover of *Science* on 28 July 1995. WGA became standard procedure for microbes: *Mycoplasma genitalium*, *Methanococcus jannaschii* and many more quickly followed.

By 1998, the Human Genome Project, financed by the NIH and the US Department of Energy (DoE), was making slow and steady progress by mapping clones, sequencing them and using the map to assemble the genome. In May of that year, everything changed. Venter announced that he and Hunkapiller (by now running Applied Biosystems) were forming a private company, Celera, to sequence the human genome more cheaply and faster than the non-commercially funded consortium, later called the International Human Genome Sequencing Consortium. They would use WGA and capillary sequencers not yet built.

The announcement, covered by the world's press, was met with incredulity (WGA would never work with a genome so complex), consternation (the human genome would be in the hands of a private company) and fear (would Congress and the Wellcome Trust cut off funding for the public project?). Days later at the Cold Spring Harbor Laboratory Genome Mapping and Sequencing meeting on Long Island, New York, the tension was palpable — and exciting. With characteristic chutzpah, Venter fanned the sparks when he suggested that the NIH and DoE teams give up on the human genome and do the mouse instead.

Jim Watson was at the meeting, asking the elite how they were to counter-attack. It was not the US cavalry but the British Grenadiers — John Sulston and Michael Morgan of the Wellcome Trust — who restored confidence. The public effort girded its loins and stepped up the pace. Interactions between public and private projects remained poisonous. Occasional attempts at reconciliation foundered on data-release issues.

Only three years later, first drafts of the human genome were published simultaneously by Celera and the public consortium. Venter shared a podium with Francis Collins, head of the US genome project, and President Bill Clinton at the White House, while UK Prime Minister Tony Blair attended via satellite link. Even this moment of triumph held no reconciliation. Venter balked at the standard requirement that all data should be provided in the paper, so Celera published in *Science*, as it had agreed, with restricted access. The public group, making all their sequence available for free, published in *Nature*. Squabbling continued over exactly how the private sequence had been assembled. Many people's worst opinions of Venter were confirmed when he admitted that most of the Celera sequence was his own, rather than that of anonymous DNA donors. One journalist called it a "high point in the annals of egotism".

Called to account

This year, Venter and Watson became the first people to have their entire genomes sequenced and made public. Both believe that deciphering our individual genetic inheritance will lead to better health: if there are no therapies yet for what is found, the risk might be minimized. Venter underlines the point throughout his book by describing what particular genes mean for him. For example, he takes a statin to lower his cholesterol levels, because he has the *E4* allele of the *APOE* gene that increases the risk of Alzheimer's disease.

I have interacted with Venter over the years since our first meeting in 1990, and have heard many strong opinions of his character. *A Life Decoded* is a fair representation of the man. It may even be more revealing than he thinks.

But the differing published accounts of the *Drosophila* and human-genome sequencing projects are reminiscent of the fable about the blind men who described an elephant by touch. Reading the books by John Sulston and Georgina Ferry (*The Common Thread: A Story of Science, Politics, Ethics and the Human Genome*), James Shreeve (*The Genome War: How Craig Venter Tried to Capture the Code of Life and Save the World*), Michael Ashburner (*Won for All: How the* Drosophila *Genome Was Sequenced*) and now Venter's contribution, it is scarcely credible that the protagonists lived through the same events. Robert Cook-Deegan's *The Gene Wars: Science, Politics, and the Human Genome* provided an authoritative, inside-the-Beltway account of the early days of the Human Genome Project, but what we need is a record of the whole project by a team of historians with no axe to grind.

Such an endeavour should begin with a comprehensive collection of material, along the lines of Thomas Kuhn's *Sources for History of Quantum Physics*. Kuhn and his colleagues interviewed the participants in, and found primary documents relating to, the greatest change in our view of the physical world since Isaac Newton. The greatest project in biology so far deserves to be similarly documented. The principals are still with us, as are their e-mails.

Chargaff called the heroes of *The Double Helix* "a new kind of scientist, one that could hardly have been thought of before science became a mass occupation, subject to, and forming part of, all the vulgarities of the communications media". Four decades on, our infinitely more vulgar media has called Venter many things: maverick, publicity hound, risk-taker, brash, controversial, genius, manic, rebellious, visionary, audacious, arrogant, feisty, determined, provocative. His autobiography shows that they are all justified. ■

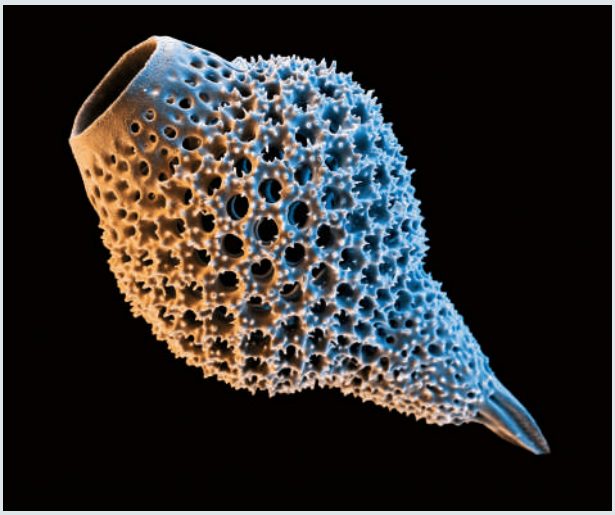
Jan Witkowski is director of the Banbury Center, New York, and professor in the Watson School of Biological Sciences at Cold Spring Harbor Laboratory. He is a co-author of *Recombinant DNA: Genes and Genomes — A Short Course*.

A gallery of micrographs

German biologist Ernst Haeckel branded radiolarians — tiny seawater plankton — one of the "art forms of nature", an accolade borne out under the scrutiny of the scanning electron microscope almost a century later. These single-celled animals come in a variety of intricate shapes, as shown in this image of a radiolarian shell. It is one from a striking collection of micrographs assembled by the Science Photo Library for *Microcosmos* by Brandon Broll (Firefly Books, 2007).

The book includes some 200 images, taken at up to 22 million times magnification, of subjects drawn from biology, mineralogy and technology. Readers can marvel at pictures of a hummingbird hawkmoth's tongue and of nanowires just ten atoms wide, of exotic gallstone crystals, butterfly scales and hairy gecko feet. Although kettle scale and a wound dressing in filigree may be a step too far into the microcosmos, there is wonder lurking in these too.

The micrographs are showcased with cunning digital artistry to impart colour. In places this borders on the fanciful (garish crystals of vitamin C and a Siberian microdiamond), but otherwise brings the pictures sharply to life.



Honest Jim talks manners

Avoid Boring People: And Other Lessons from a Life in Science
by James D. Watson
Oxford University Press/Alfred Knopf: 2007. 362/256 pp. £14.99/$26.95

Huntington F. Willard

The opening line of James D. Watson's *The Double Helix* — "I have never seen Francis Crick in a modest mood" — has become part of scientific literary folklore. As I re-read my dog-eared and musty copy of his 1968 kiss-and-tell about the discovery of the structure of DNA, I was more drawn to his closing line: "I was twenty-five and too old to be unusual." Now, decades later, Watson has apparently decided that he is no longer too old.

His latest book, a memoir with the playfully ambiguous title *Avoid Boring People*, weaves a deliciously detailed account of his life both in and out of science with a series of lessons drawn from those experiences. Each chapter ends with a homily (the title being but one) recounted in an appendix that serves as a sort of '108 commandments' to younger scientists. Some are silly, some are off the mark after so many decades, and some are painful to acknowledge. But most are insightful, useful and on target about science, competition, leadership, teaching and academic success.

The book runs from Watson's childhood in Chicago, through largely familiar accounts of his years in graduate school and then in Cambridge, to tales of political intrigue at Harvard, and finally to his decision to head the Cold Spring Harbor Laboratory in New York. Each chapter title has a manners theme ("Manners learned…", "Manners followed…", "Manners demanded…", and so on), an ironic twist for those who have witnessed the author in high dudgeon. (Only Watson's advice on manners could include an injunction "to be forthright and call crap 'crap'".)

His remarkable recall of events, both important and trivial, gives the reader the feeling of being there. He first used this technique in *The Double Helix*, a book he originally wanted to call *Honest Jim*. The details, in interwoven paragraphs that alternate between academic science and his social escapades, reveal conflicting moments of deep insecurity and robust confidence in "using my head".

The book is full of insight into Watson and into a life in science. My favourite is his reading of *Arrowsmith* by Sinclair Lewis, the tale of a young man who enters science to save the world from cholera. Watson discovered the book as a teenager and was still assigning it to his biochemistry classes at Harvard some 30 years later. With pride, he recalls one colleague comparing *The Double Helix* to the earlier classic.

We learn who and what has earned Watson's respect, affection and tenderness: his father, his wife Liz, the University of Chicago, former Chicago president Robert Hutchins, teaching, Harvard students, art, and those he injudiciously refers to as 'girls'. And also what earned his ire, jealousy or disregard: episcopalian Republicans, religion, Harvard, former Harvard president Nathan Pusey, and athletes. He is at all times blatantly but entertainingly honest about his likes and dislikes.

HARVARD NEWS OFFICE/WATSON COLLECTION

Gossip, manners and passions shape Watson's mischievous memoirs, but some tales are yet to be told.

Sadly, and without explanation, the book ends too soon, at the point where Watson leaves Harvard in 1976. I was ready for another 30 years of tales untold — about his years at Cold Spring Harbor Laboratory, being the first leader of the US Human Genome Project, runins with Craig Venter and with the National Institutes of Health director Bernadine Healy, and deciding to sequence his own genome. There is no reflection on the bookends of a life in science that has spanned just over half a century — at one end standing beside the first, crude model of the double helix and at the other contemplating his own DNA sequence. There's another book yet to be written to complete Honest Jim's story.

One revealing part of the book is an odd epilogue, odd because it focuses not on events in Watson's but in Harvard's life. Watson is highly critical of science at Harvard, while expressing sympathy for the demise of former Harvard president Larry Summers. These events would seem to be largely irrelevant to the rest of the book, had Watson not been in hot water in the mid-1970s over 'girls' in science, and had he not been curious about the role of the genome in shaping human intellectual ability and in predisposing to such 'developmental failures' as autism, schizophrenia and Asperger's syndrome. He tellingly concludes: "If Summers' tactlessness does, in fact, have a genetic basis, much of the anger toward him should rightly yield to sympathy." In genome, *veritas*.

Crick once famously complained that *The Double Helix* presented "the history of science as gossip"; *Avoid Boring People* might be viewed as the history of gossip presented as science. Lessons can still be drawn from gossip by those who have known Watson or by those who are the poorer for never having seen him in action. Unusual or not, Watson remains one of the most fascinating scientists of our time, as iconic in some respects as his double helix. ■

Huntington F. Willard is director of the Duke Institute for Genome Sciences & Policy at Duke University in Durham, North Carolina 27708, USA.

From alchemy to chemistry

Andreas Libavius and the Transformation of Alchemy: Separating Chemical Cultures with Polemical Fire
by Bruce T. Moran
Science History Publications: 2007. 344 pp. $49.95

Chymists and Chymistry: Studies in the History of Alchemy and Early Modern Chemistry
edited by Lawrence M. Principe
Chemical Heritage Foundation/Science History Publications: 2007. 274 pp. $45

Philip Ball

The Communist Party of Britain (Marxist–Leninist), it has been said, had no scorn for right-wingers to match that it reserved for the parenthesis-free Communist Party of Britain. To judge from the writings of the German chymist Andreas Libavius, little has changed. The invective he hurled at, and received from, contemporaries whose views we now struggle to distinguish, is a sober reminder of how historical distance turns violent academic spats into hair-splitting.

Libavius (c.1550–1616) has always been a marginal figure in histories of chemistry. He usually features as the author of *Alchemia* (1597), commonly called the first textbook of chemistry. The book served as the model for Jean Beguin's *Tyrocinium Chymicum* (1610) and Nicolas Lémery's *Cours de Chimie* (1675), in which one can start to see something like early modern chemistry taking shape.

As Bruce Moran argues in *Andreas Libavius and the Transformation of Alchemy*, his majestic survey of Libavius's life and work, *Alchemia* is by no means the key to Libavius's role in chemical history. By awarding it a place in chemistry's lineage, historians have imposed some apparent order on the emergence of the subject from the sooty fumes of alchemy. Moran makes a convincing case that this is an artificial narrative. The transition was unruly — in all respects — and mocks attempts to label its players as progressives or conservatives.

Moran's scholarly efforts are all the more valuable because they could seem in some ways unrewarding. Libavius was not an especially innovative thinker. He spent little time in the laboratory and what he wrote on chymistry — the transitional discipline between alchemy and chemistry — was, he admitted, taken mostly from other sources. He expended much energy in furious disputes, involving the kind of bitter, frequently scatological diatribes that, while amusing at first, soon become tiresome. "You will never leave the contest unless you have left behind a barb," wrote the Parisian doctor Jean Riolan in 1606 of Libavius. All this, Moran admits, can make Libavius seem "more like an off-putting sour-puss than a compelling or attractive historical figure".

Yet in these battles Libavius reveals a great deal about the issues — philosophical, political and religious — on which the establishment of chemistry as an academic discipline hinged.

Libavius's sharp tongue recalls that of the figure whose writings triggered much of the argument: the Swiss alchemical physician Paracelsus (1493–1541). Medicine by the end of the sixteenth century is often depicted as split into two camps: the traditionalists, who relied on the humoral theory of Hippocrates and Galen, and the chemical physicians ('iatrochemists') who used Paracelsian cures. This is too simple a picture, Moran shows. There were "Hippocratic hermeticists, Galenochemists, natural and hermetic chymiatrists" and others, all denouncing one another venomously.

Alchemy, depicted here by van der Straet, begat the study of chymistry.

ALFREDO DAGLI ORTI/PALAZZO VECCHIO FLORENCE/ART ARCHIVE

Libavius had to be pugnacious because he didn't really fit into any camp. He despised Paracelsians, but was highly critical of the traditionalists and came to the defence of the Paracelsian court physicians to Henry IV of France when they were attacked by the Galenist Parisian medical faculty. He seems almost to have courted enemies by taking positions that were superficially contradictory. He wanted chymistry to be considered a serious liberal art, yet called it "a bilge-flood and chaos of impurity and human dregs". His damning assessment was, however, aimed at chymistry as he saw it being done, polluted by infernal Paracelsians. He thought it should be a science that rejected mystical speculation and combined manual arts learnt through experience and philosophical principles derived from Aristotle.

The book isn't an introductory read. You could be misled, for example, if you do not already know that by 'Ramon Lull' and 'Roger Bacon' Moran means a body of writings attributed, often apocryphally, to those authors, or that Basil Valentine is probably a pseudonym of a late-sixteenth-century writer. And he takes it for granted that the reader will recognize Libavius's portrayal of Paracelsianism to be mostly a caricature. But the book opens up a neglected period that chemical historians have been too eager to skate over in the rush to get from Paracelsus to Robert Boyle.

Indeed, the role of alchemy in the genesis of chemistry has become a hot topic, as Lawrence Principe says in his introduction to *Chymists and Chymistry*. Time was when those who studied it risked being thought of as lacking in judgement, or worse. Several recent books, such as William Newman's study of Daniel Sennert's atomistic theories, *Atoms and Alchemy*, have shown that 'modern' theories of matter and chemical change actually grew from alchemy, rather than supplanting it.

Chymists and Chymistry is the product of a major conference, hosted by the Chemical Heritage Foundation in Philadelphia in 2006; it brings together papers with an immense scope, straddling the eras of Paracelsus and Antoine Lavoisier. Particularly welcome are contributions on archaeological analyses of alchemical equipment — a neglected arena in a subject usually reliant on text alone. As one of the contributors, Marcos Martinón-Torres of University College London, observes, pioneer historians of alchemy were often chemists who could be historically naïve; the danger now is that the subject be conducted only by historians who ignore materials science. Libavius was right in that respect: we need both books and benchtops. ■

Philip Ball is a consultant editor at *Nature*. His books include *The Devil's Doctor: Paracelsus and the World of Renaissance Magic and Science*.

SCIENCE IN CULTURE

Paris gets a new cultural crucible

Scientific breakthroughs are reached through aesthetic, as well as scientific methods, argues a bioengineer, who is this week opening a culture centre to explore such creativity.

David Edwards

In the early 1990s, I worked with Howard Brenner at the Massachusetts Institute of Technology (MIT) on problems of fluid mechanics. We prided ourselves on a certain way of seeing truth in form. "Aesthetics" was how Brenner taught me to describe what we were up to.

My subsequent career showed me that, although science and art do not seek anything near the same ends or generally the same means, the two become one at pioneering moments. We may use the scientific method, of deductive experiment and analysis, to pursue our ideas, but those ideas are born and propelled in new directions through the aesthetic method — moments of induction and association, when our sense of beauty tells us what ought to be true. We innovate in science by viewing it as art.

In the mid-1990s, while teaching at Pennsylvania State University, I collaborated with MIT materials scientist Robert Langer. We needed to create an insulin particle for the treatment of diabetes that would effectively fly through the air and land in the lungs. It struck me that we should redesign the particles to resemble kids' whiffle balls, perforated plastic spheres designed for indoor baseball. Like many new ideas, this one came about not through deduction but by association, in this case of a drug particle and a childhood memory. My idea arrived through the aesthetic method. We put the idea to the test with laboratory experiments — the scientific method — and published it in *Science*.

Ironically, such an approach seems particularly necessary to commercializing science. Interacting with bankers and entrepreneurs showed me how intuition and risk-taking are central to the culture of business. On the basis of such experiences, I started the Idea Translation Lab (ITL) at Harvard University. Here, students from all disciplines develop ideas ranging from a new application of microfluidics in architecture to music as medicine. ITL student teams helped in the start-up of the Massachusetts-based pharmaceutical company Pulmatrix, and the not-for-profit infectious disease organization Medicine in Need, or MEND, based today in the United States, France and South Africa.

Unfortunately, most academic, humanitarian, cultural and industrial organizations discourage the simultaneous development of aesthetic and scientific methods. Specialization is as essential to science as it is to art. Yet without the freedom, or the time, to wander in our research, to explore avenues of inquiry that cross conventional boundaries, we jeopardize something elemental to the pioneering spirit.

A. POTIGNON

Le Laboratoire will showcase experiments that fuse art and science.

So I am creating the cultural centre Le Laboratoire, opening this week in Paris, with the proceeds from the sale of the company founded to develop the whiffle-ball idea. Le Laboratoire will present to the public art and design works-in-progress resulting from seasonal experiments. Leading international artists and scientists will explore, for example, the experience of cell division through visual art, making plants smarter using fluid mechanics, personalizing urban homes through biometric testing, or designing a synthetic world where avatars evolve through the uploading of personal biometric data.

Such experiments will aim to stimulate creativity with cultural, commercial, social and educational ramifications. Initial partners include Apple, Société Générale, the Wellcome Trust, the French National Research Agency for AIDS, Connaissance des Arts, Radio Classique and Harvard's ITL.

Already, an international constellation of cultural organizations fosters art and science crosstalk. The 40-year-old *Leonardo* journal, and the more recent organization Art & Science Collaborations, document and encourage interdisciplinary partnerships. Many science laboratories, museums and businesses host artist residences, including the University of California in Berkeley, the Natural History Museum in London and Xerox in Silicon Valley, California.

Our fellow art-and-science exhibition spaces include the Science Gallery due to open in 2008 at Trinity College, Dublin, and the new Wellcome Collection galleries in London, where biomedical science is explored through artistic expression. The latter is funded by the Wellcome Trust, long a major donor to polymathic collaborations.

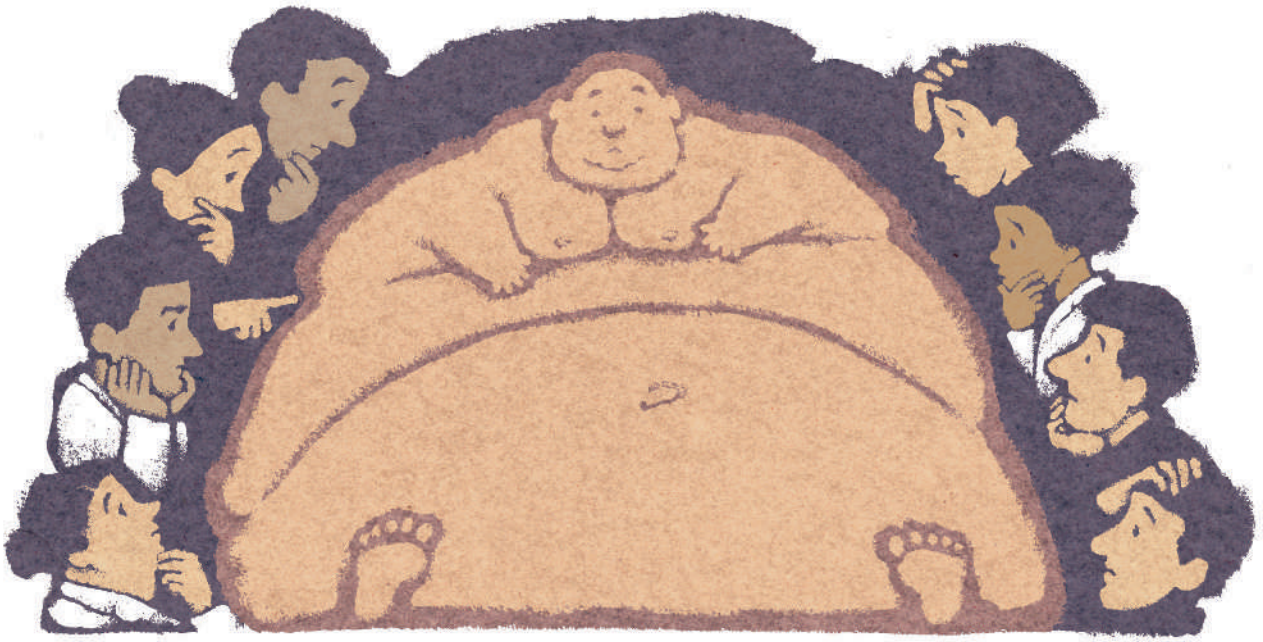
But Le Laboratoire is the first experiment-driven art and science incubator. The centre aims to nurture and showcase a potent creative process that merges what we conventionally refer to as art and science. This method of human inquiry is at once intuitive and deductive, sensual and analytical, directed and comfortable with uncertainty, embracing nature in its complexity and able to model nature in its essence. ■

David Edwards is a professor of bioengineering at Harvard University, 29 Oxford Street, Cambridge, Massachusetts 02138, USA. He is the author of the forthcoming books *Niche* and *ArtScience: Creativity in the Post-Google Generation*.

Le Laboratoire opens in Paris on 19 October www.lelaboratoire.org

Big lessons for a healthy future

This week's report on obesity policy in the United Kingdom highlights three challenges for scientists and politicians working together.



D. PARKINS

David A. King and Sandy M. Thomas

Science today is both a driver of technological innovation and an important resource for shaping public policy. In the United Kingdom, growing recognition of how science can contribute to health, wellbeing and the economy, has led to the appointment of six departmental chief scientific advisers in seven years, and improvements in the way that government departments use scientific analysis in policy-making. Countries such as Canada, France and Germany have taken similar steps.

Problems remain, though, in the relationship between science and governance. These are illustrated by the latest report from the UK government's new Foresight programme, which was established by David King (www.foresight.gov.uk). Published this week, the results from the two-year study of obesity improve our understanding of the causes of the worldwide epidemic. They also highlight three key areas of potential conflict between the worlds of science and policy-making.

First, individual government departments struggle with multi-faceted problems; second, political lifetimes are short compared with interventions that need several decades to have an impact; and third, problems bedevilled by uncertainty can be marginalized by policy-makers eager for certainty. Our experiences with the project suggest that strategic cross-department policy tools are needed to improve coordination between scientists and policy-makers.

The Foresight programme was born out of the success of the scientific contribution to tackling the country's 2001 foot-and-mouth epidemic. To provide long-term strategic advice, Foresight projects draw together research to study a range of possible futures with different social, economic and political contexts. They are designed to help policy-makers anticipate risks and opportunities, test possible interventions and foresee possible outcomes.

"Ministers want straight-forward solutions, but our analysis shows that no one policy will fix the obesity problem."

One of the most important findings of the Foresight obesity project is that individuals have much less choice in the matter of their weight than we may often assume. Our analysis shows that the current epidemic of obesity does not arise from individual over-indulgence or laziness. Instead, human biology has become out of step with the structure of society.

We evolved to respond to hunger by eating; we are only weakly able to notice, and stop when we have had enough. This was an effective survival strategy in prehistoric times when food was scarce. Now high-energy, cheap foods abound, as do labour-saving devices, motorized transport, sedentary work and the association of eating with indulgence. These conspire to create an 'obesogenic' environment. The increasing prevalence of obesity is a consequence of modern life.

To stay trim, most people would need to pay an overwhelming amount of attention to overriding their instincts, habits, conflicting aspirations, upbringing, peers and the environmental cues in their day-to-day lives. Hence fewer and fewer people manage it. Our study suggests that by 2050, in the United Kingdom about 60% of men, 50% of women and 25% of children will be obese. The associated chronic health problems are projected to cost society an additional £45.5 billion (US$93 million) a year.

Politicians and the public would like to avoid this alarming scenario. The causes of obesity are relatively well understood — the challenge lies in how to develop effective, evidence-based policy interventions.

Many hands make light work

The first problem is that ministers want straight-forward solutions. Our analysis shows that no one policy will fix the obesity problem. We modelled how physiological, technological, economic and cultural factors fit together and how a range of changes would affect this complex system. Lone interventions such as a new appetite-suppressing drug could well worsen the matter.

NATURE|Vol 449|18 October 2007

Any response has to be multifaceted, or it will almost certainly fail. We need policies aimed at different life stages, in particular early on to establish appropriate child growth, healthy eating and activity habits. We also need to increase everyday activity levels through the design of the built environment and transport systems, and to work out how to shift consumer purchasing patterns to favour healthy options. The promise of community-level interventions is hinted at by the Fleurbaix-Laventie City Health study in France.

Dealing with obesity will require intense cooperation across many government departments. The UK government, like many others, operates through a strongly vertical management structure of separate ministries, with few mechanisms or incentives for horizontal collaboration. For a problem such as obesity, this can result in a 'policy cacophony' of conflicting approaches.

To reduce such problems, the Foresight project involved, from the start, ministers from the Departments of Health, Culture, Media and Sport, and Children, Schools and Families, alongside food retailers, local government and sports associations, and other stakeholders who will be responsible for implementing the resulting policy initiatives.

In managing complex issues, structured comprehensive collaboration such as this must become the norm. A useful example is the UK Office of Climate Change, which involves six government departments. It was established because global warming was high on the agenda of the prime minister, the treasury and parliament, and there was wide recognition of the need for collaboration across normal departmental boundaries.

Indeed there are strong parallels between climate change and obesity. Companies in the energy and food sectors generally encourage people to consume more. In both domains, failure to agree and act now will lead to serious adverse consequences in just a few decades — because reversal of the trends may cease to be an option in both cases. And, as for climate change, communication is central to changing the obesity landscape. Our 'system map' for the different drivers of obesity is so intricate that the clearest possible explanations are needed for policy-makers and society. For scientific advice to underpin government action, communications skills must be a much bigger part of scientific training and culture.

The long view

The second problem for science policy-making is timescale. Brief political lifetimes mean policy-makers attach great value to rapid results. But in the case of obesity, solutions will need to be implemented over several decades to have any impact.

For such problems, it might be necessary to change the way that policy is made. For instance, the Ministry of Defence has traditionally not been subjected to the shorter-term goals given to other departments, since it is widely acknowledged that national security transcends day-to-day politics. The Committee on Climate Change is another useful example, as proposed in the United Kingdom's forthcoming climate-change bill. The committee will operate at arms length from the government, akin to the Bank of England, with access to levers that can drive policy over decades, with considerably more power and responsibility to parliament than a conventional government task force.

The third friction illustrated by our obesity work concerns attitudes to uncertainty. Ministers want clear messages but scientists are uncomfortable speculating. And analysis of complex problems often reveals high levels of uncertainty. Unfortunately, nobody yet knows exactly which policy levers will slim the nation. Most research into obesity has focused on its causes, rather than on the effectiveness of potential interventions. More cross-disciplinary research over long time periods is needed.

The Foresight project was designed to help both scientists and policy-makers embrace this uncertainty. Through a combination of systems modelling and scenarios work, its outputs can help to devise an initial strategy. This strategy will then need to be monitored over different timescales, and modified as it becomes clearer which actions are the most effective.

This approach, which we term 'practice-based evidence', is revolutionary. It means that scientists can no longer simply hand over the results of their research and leave the rest to politicians. They must be fully engaged in the assessment and refocusing of strategies. They must also understand the constraints on politicians, for instance the importance that any course of action is acceptable to the public. Politicians should encourage scientists to be involved in ongoing policy processes, and accept that one can learn even from initiatives that fail.

Mutual understanding, such as that fostered by the new Foresight programme, is the foundation of all the best relationships. Both sides learn to appreciate their respective abilities and constraints. For instance, to be effective advisers rather than simply commentators, scientists have to take more responsibility for the wider implications of their research. To make good use of that advice, politicians have to realize that although science will not necessarily give the answers they want, it can give the best available analysis for moving forward. ■

David A. King is the UK Government chief scientific adviser and head of the Government Office for Science. Sandy M. Thomas is director of the UK government's Foresight programme.

For more essays and information see http://nature.com/nature/focus/scipol/index.html.

NEWS & VIEWS

PALAEOANTHROPOLOGY

The coast in colour

Sally McBrearty and Chris Stringer

A South African cave overlooking the Indian Ocean was apparently a desirable residence for early humans. The site has provided rich evidence for the early use of colour and marine resources.

On the basis of both fossil and genetic data, we know that *Homo sapiens* had evolved in Africa by 150,000–200,000 years ago. But the time and manner of human behavioural evolution are less clear. Exploiting marine resources, producing complex technology and manipulating symbols are all symptomatic of modern human activity. When did such behaviours appear, and how did the process relate to the morphological evolution of our species?

On page 905 of this issue, Marean *et al.*[1] provide strong evidence that early humans displayed key elements of modern behaviour as far back as 165,000 years ago at Pinnacle Point on the coast of South Africa (Fig. 1). The evidence is in the form of shellfish, haematite (red ochre) used as a colouring agent, and small stone 'bladelets'. The site provides a rare glimpse into human adaptation to coastal conditions during a time for which most evidence elsewhere has been scuttled by subsequent rises in sea level.

The earliest unequivocal fossils of *H. sapiens* are crania from Ethiopian sites at Omo Kibish and Herto, respectively dated to about 195,000 and 160,000 years ago. Genetic estimates[2] for the origin of our species also lie in the interval 100,000–200,000 years ago. But the behavioural repertoire and precise geographical range of these early human populations in Africa are unevenly documented.

There are two distinct views about the relationship between anatomical and behavioural evolution in early *H. sapiens*. Some workers favour a late and rather sudden origin for behavioural modernity, dating to around 45,000 years ago at the transition from the Middle Stone Age to the Later Stone Age[3]. In this view, behavioural change lagged considerably behind anatomical change, and may have resulted from a sudden neurological shift. A competing interpretation is that beads, art objects and other forms of technological and behavioural complexity emerged gradually over the course of the Middle Stone Age (some 285,000 to 45,000 years ago), tracking morphological evolution more closely[4]. In this view, early *H. sapiens* were essentially neurologically and cognitively identical to modern humans, and new behaviours seen in the archaeological record resulted from human innovation, sometimes in response to the pressures of population growth or environmental change.

The ability to manipulate symbols is

C. W. MAREAN

Figure 1 | A home at Pinnacle Point. Marean and colleagues' evidence[1] for modern behaviour in early humans, dating to 165,000 years ago, comes from site PP13B, the cave with the walkway entering it. Many coastal archaeology sites elsewhere must have been swept clean by sea-level rises during interglacials. The materials at Pinnacle Point were spared such a fate because of the site's elevation.

considered an essential part of modern human cognition and behaviour[5], although definite traces of symbols in the archaeological record are difficult to recognize and are often obscured by the ravages of time. All humans today express their social status and group identity through visual clues such as clothing, jewellery, cosmetics and hairstyle. Shell beads, and haematite used as pigment, show that this behaviour dates to 80,000 years ago in coastal North and South Africa[6,7], and to perhaps 110,000 years ago in western Asia[8]; and there are even earlier records of microliths and pigment use in Africa (Fig. 2).

Marean *et al.*[1] now describe 57 pieces of haematite, many apparently ground for use as a colouring agent, from cave PP13B at Pinnacle Point. Like haematite from the nearby site of Blombos, South Africa[6], some of the pieces of haematite from Pinnacle Point are incised, either as an aid to grinding, or perhaps as decoration, or even as elements of a notational system. The Pinnacle Point evidence is significant because it suggests that early humans in Africa inhabited a cognitive world enriched by symbols before 160,000 years ago.

Might the haematite have been used instead for some utilitarian purpose? Experimental replication demonstrates that ochre has little of its claimed utility as a hide preservative[9]. Archaeological and ethnographic evidence shows, however, that ground haematite was added to adhesives used to attach stone artefacts to bone or wooden shafts, and experimental replication demonstrates that ochre improves the durability and workability of the mastic[10]. But the consistent selection of the most brilliant reds for use in the adhesive medium by the inhabitants of Pinnacle Point, and of other African sites dating to the Middle Stone Age[1,9], cannot be so easily explained. Ochre seems to have been a material with both symbolic and utilitarian functions. The colour red is fundamental to colour classifications in all known human societies[11], and it seems probable that the substance was indeed used for body painting and to colour artefacts by 165,000 years ago.

The presence of stone bladelets at Pinnacle Point may also be significant. Miniature stone tools were important to African technology after 40,000 years ago, when small geometric implements were mounted, often in multiples, as projectiles. These no doubt gave their wielders an advantage over populations limited to hand weapons. Microlithic technology is known from South and East Africa at around 70,000 years ago, but the Pinnacle Point bladelets are nearly 100,000 years older. However, they show no sign of shaping into geometric tools, and the fact that they comprise the small end of a continuous size distribution for blades at the site indicates that they may not have been deliberately designed[12]. Microwear and residue analysis might reveal if and how the bladelets were hafted and used.

At about 165,000 years old, the remains of molluscs at Pinnacle Point are the earliest securely dated evidence for the systematic exploitation of shellfish as food. Marean *et al.*[1] suggest that shellfish may have been seen by the Pinnacle Point people as a 'famine food', to be consumed when more preferred items were scarce during the challenging cold and dry conditions that pertained in Africa between about 195,000 and 130,000 years ago. Neanderthals are known to have cooked shellfish in caves in Italy as early as 110,000 years ago[13], and shellfish may well have been exploited routinely by even earlier coastal populations. However, many cave repositories around the world were swept clean by the rise in sea level during the last interglacial, which ended around 115,000 years ago, and others have been submerged by the sea-level rise of the present interglacial.

Anthropogenic deposits fortunately escaped this fate at Pinnacle Point owing to its elevation. By comparing dates obtained from uranium-isotope and luminescence dating with bathymetric records and reconstruction of topography based on geographic-information-system data, the authors show that the coastline lay within reasonable foraging distance of the site only at around 167,000 years ago, in good agreement with the key dated phase of human occupation. Better understanding of the ancient topography may help in discovering other sites in the vicinity of Pinnacle Point, and advances in dating techniques may allow further evidence from this critical interval to be identified elsewhere on the continent, in both coastal and inland settings. We should expect surprises as the evidence for behavioural complexity in early humans continues to accumulate. ■

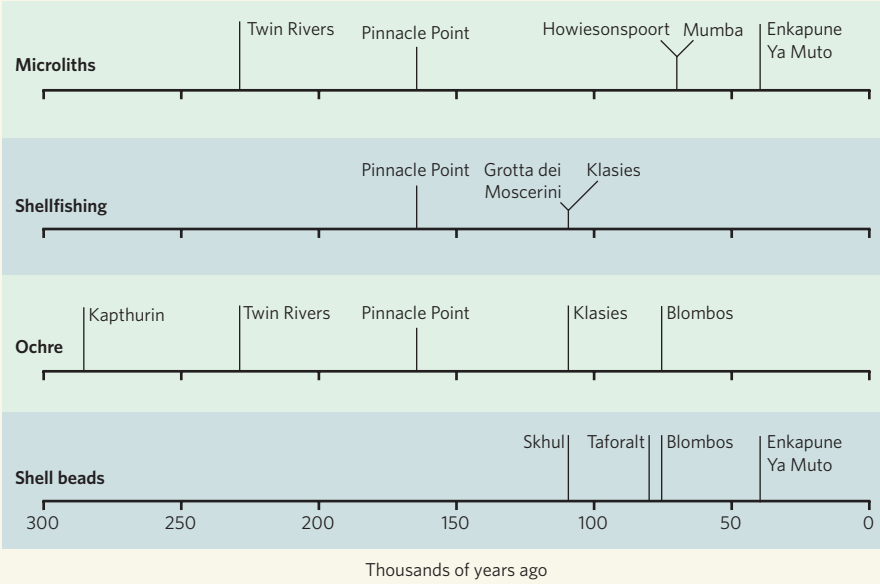


Figure 2 | Timeline of some notable archaeological evidence for modern behaviour in early *Homo sapiens*. For context, the earliest unequivocal fossils of *H. sapiens* come from two sites in Ethiopia, and date to about 195,000 and 160,000 years ago. The behavioural evidence is in the form of the production of microliths (very small stone flakes); the exploitation of shellfish as a food resource; the use of ochre (haematite) as a pigment; and the creation of shell beads. The coastal site of Pinnacle Point, South Africa, has yielded evidence for the first three of these activities, as described by Marean *et al.*[1]. Site locations: Twin Rivers, Zambia; Howiesonspoort, South Africa; Mumba, Tanzania; Enkapune Ya Muto, Kenya; Klasies, South Africa; Kapthurin, Kenya; Blombos, South Africa; Skhul, Israel; and Taforalt, Morocco. Grotta dei Moscerini, in Italy, provides an early example of Neanderthal exploitation of marine resources.

Sally McBrearty is in the Department of Anthropology, University of Connecticut, Storrs, Connecticut 06269-2176, USA.
Chris Stringer is in the Department of Palaeontology, The Natural History Museum, London SW7 5BD, UK.
e-mails: mcbrearty@uconn.edu; c.stringer@nhm.ac.uk

1. Marean, C. W. *et al. Nature* **449,** 905–908 (2007).
2. Stringer, C. in *The Prehistory of Africa* (ed. Soodyall, H.) 10–20 (Ball, Johannesburg, 2006).
3. Klein, R. G. & Edgar, B. *The Dawn of Human Culture* (Nevraumont, New York, 2002).
4. McBrearty, S. & Brooks, A. *J. Hum. Evol.* **39,** 453–563 (2000).
5. Henshilwood, C. S. & Marean, C. W. *Curr. Anthropol.* **44,** 641–642 (2003).
6. Henshilwood, C. S. in *Rethinking the Human Revolution* (eds Mellars, P. & Boyle, K.) (MacDonald Institute, Cambridge, in the press).
7. Bouzouggar, A. *et al. Proc. Natl Acad. Sci. USA* **104,** 9964–9969 (2007).
8. Vanhaeren, M. *et al. Science* **312,** 1785–1788 (2006).
9. Watts, I. *S. Afr. Archaeol. Bull.* **57,** 1–14 (2002).
10. Wadley, L., Williamson, B. & Lombard, M. *Antiquity* **78,** 661–675 (2004).
11. Kay, P., Berlin, B., Maffi, M. & Merrifield, W. in *Color Categories in Thought and Language* (eds Hardin, C. & Maffi, I.) 21–58 (Cambridge Univ. Press, 1997).
12. Ambrose, S. H. in *Thinking Small: Global Perspectives on Microlithization* (eds Elston, R. G. & Kuhn, S. H.) 9–29 (Am. Anthropol. Assoc., Washington DC, 2002).
13. Stiner, M. C. *Honor Among Thieves* (Princeton Univ. Press, 1994).

EARTH SCIENCE

An Indian cheetah

R. Dietmar Müller

After the supercontinent of Gondwanaland broke up, the part that became India diverged especially swiftly from the other fragments. The explanation for this might lie in the loss of India's deep roots.

Look up 'speed boosting' on the Internet and you'll find recipes for boosting the speed of computers, modems, cars, photographic film, gas turbines and even your golf cart. But how would you increase the speed of a continent ploughing through Earth's viscous, churning mantle? Kumar *et al.* (page 894 of this issue[1]) have come up with an answer to precisely this question. The recipe is simple: remove the continent's deep roots by heating from below. Given that older continental material tends to be thicker, such treatment will make an ageing continent look younger. It could also transform it from a sloth to a cheetah in terms of its response to tectonic forces, allowing it to race along at speeds of up to 200 mm per year.

Earth's surface is composed of an interlocking mosaic of tectonic plates that consist of crust and a portion of the upper mantle, together known as lithosphere. Some plates are entirely composed of oceanic lithosphere, created by seafloor spreading at mid-ocean ridges. However, most plates carry continents around with them, many of which have a history reaching back to the early Archaean some 3,800 million years ago. Tom Jordan[2] was the first to recognize that the mantle beneath young continental crust is fundamentally different from that beneath old continental crust, with older crust being underlain by thick roots or keels of relatively light and buoyant material. Because of its buoyancy, this lithospheric mantle material cannot be easily removed. It is now well accepted that many cratons — the term for old and stable parts of continental crust — are underlain by roots that may be more than 250 km thick.

The realization that old continents have giant lithospheric keels reaching deep into the mantle spurred a new controversy: do these keels act to slow down or speed up the movement of the continents above them? One view[3] is that they speed continents up, allowing the convecting mantle to drag them along more effectively. The opposite argument[4], based on combined geodynamic modelling and palaeomagnetic data, is that speed increases with decreasing root depth for plates being driven by tectonic forces closer to the surface. Kumar and colleagues' results[1] provide strong evidence that the latter premise is correct.

Using a recently developed seismic technique, the 'S-wave receiver function', Kumar *et al.* measured the lithospheric thickness of several continents with unprecedented accuracy. The continents concerned — India, Africa, Antarctica and Australia — were all part of Gondwanaland, the supercontinent that broke up with the creation of the Indian Ocean. The authors show that cratons in South Africa, Antarctica and Australia are more than 180 km thick, whereas Indian lithosphere is only about 100 km thick, even where the crust is Archaean in age (see Fig. 4 of the paper on page 896). Diamond-bearing kimberlite, a rock type produced from melts arising from considerable depth, demonstrates that before Gondwanaland's break-up India's lithosphere must have been much thicker than it is today.

Kumar *et al.*[1] argue that the most likely time for India to have lost its continental roots is when a large upwelling of especially hot rock — a mantle plume — hit Gondwanaland during or immediately after its break-up, although the exact timing is uncertain. Their contention is that the plume both broke up Gondwanaland and melted India's roots but not those of the other fragments.

The most intriguing part of Kumar and colleagues' analysis is that there seems to be a clear connection between India's lost lithospheric roots and its exceptionally fast northward motion from about 130 million years ago until about 50 million years ago. That motion is well documented from palaeomagnetic data and rates of seafloor spreading, and sets India apart from all its neighbours. Africa, Antarctica and Australia all have thick lithospheric roots and have all moved relatively sluggishly, at about 20–40 mm per year, during that same time interval.

The effect of India's superfast plate motion on seafloor-spreading rates is illustrated in Figure 1. Extremely fast seafloor-spreading rates (half-rates of more than 70 mm per year) are typically found where plates without continents are diverging, as in the Pacific Ocean. In the Indian Ocean such superfast rates are unexpected, but clearly evident, underlining India's unique plate-tectonic behaviour among continent-bearing plates.

But what about Gondwanaland's previous life? On the basis of apparent polar-wander curves derived from palaeomagnetic data, drift rates of the supercontinent may have exceeded 180 mm per year during parts of the Devonian and Carboniferous periods (about 400 million to 300 million years ago), suggesting that continents have overcome tectonic speed limits before[5]. Does that mean that there are other circumstances that allow plates with large continents to behave like a Ferrari? As high-quality palaeomagnetic databases grow, continental drift will be mapped more accurately. Integrated with other geological data constraining lithospheric thickness through time and space, plate-tectonic reconstructions provide a continental speedometer for unravelling the fundamental causes of speed limits.

Kumar and colleagues' seismic observations and maps of lithospheric thickness will

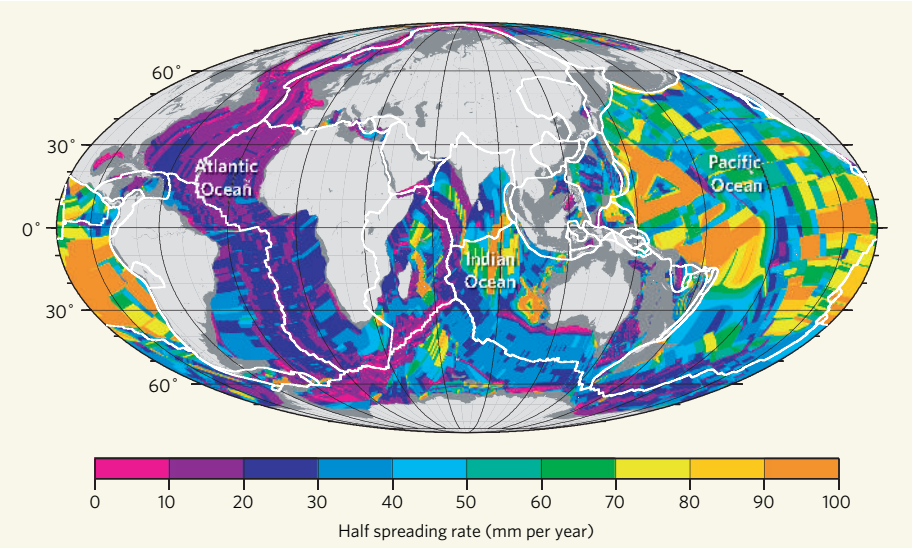


Figure 1 | Global seafloor-spreading rates of individual mid-ocean-ridge flanks (called half-rates). Plate boundaries are delineated by white lines; continental margins are medium grey and continents light grey. Magenta and blue indicate the slow spreading typical of plates that carry large continents; green, yellow and orange represent the fast spreading typical of purely oceanic plates in the Pacific. The yellow–orange patches seen on the floor of the Indian Ocean are an exception created by a super-fast spreading regime. They represent ocean floor created during the fast motion of India away from Africa, Antarctica and Australia (other components of the supercontinent Gondwanaland), which Kumar *et al.*[1] attribute to the loss of India's deep lithospheric roots. (Figure based on a revised data set from ref. 8.)

probably prove less controversial than their interpretations of the associated geodynamic events. For example, the idea that the break-up of Gondwanaland was caused by a large mantle plume is contentious. Data on the ages of certain circum-Indian Ocean basalt rocks and of the oldest Indian Ocean sea floor suggest that the split of India from Australia and Antarctica occurred about 130 million years ago. With one small exception, that precedes evidence of plume activity, in the form of large-scale emplacements of igneous rocks, by at least 10 million years[6]. The assumption that large-scale, sub-continental melting is necessarily caused by mantle plumes has also been challenged[7], with the suggestion that continental aggregation may promote such melting without the involvement of a plume. Even though the timing and circumstances of the loss of India's lithospheric roots will remain controversial, one thing is sure: Archaean lithosphere is not for ever. ■

R. Dietmar Müller is in the School of Geosciences, University of Sydney, NSW 2006, Australia.
e-mail: dietmar@geosci.usyd.edu.au

1. Kumar, P. *et al. Nature* **449,** 894–897 (2007).
2. Jordan, T. H. *Nature* **274,** 544–548 (1978).
3. Stoddard, P. R. & Abbott, D. *J. Geophys. Res.* **101,** 5425–5434 (1996).
4. Gurnis, M. & Torsvik, T. H. *Geology* **22,** 1023–1026 (1994).
5. Meert, J. G. *et al. Nature* **363,** 216–217 (1993).
6. Coffin, M. F. *et al. J. Petrol.* **43,** 1121–1137 (2002).
7. Coltice, N., Phillips, B. R., Bertrand, H., Ricard, Y. & Rey, P. *Geology* **35,** 391–394 (2007).
8. Müller, R. D., Roest, W. R., Royer, J.-Y., Gahagan, L. M. & Sclater, J. G. *J. Geophys. Res.* **102,** 3211–3214 (1997).

STRING THEORY

Back to basics

Hermann Nicolai

Long touted as a theory of everything, it seems that string theory may at last succeed as a theory of something very specific — the interactions of particles under the strong nuclear force.

Whether string theory can live up to its claim of being a 'theory of everything', and whether it will ever produce a falsifiable prediction as such, remain hotly debated questions. Meanwhile, developments in a quieter side-alley[1–8] indicate that the theory might be about to deliver something of its original promise: helping us to understand the physics of interactions mediated by the strong nuclear force.

String theory was born in the 1960s, when physicists tried to put order into a bewildering wealth of phenomena appearing at subnuclear distance scales. Here, the strong interaction dominates the other three fundamental forces of nature: gravity, electromagnetism and the weak nuclear interaction. Scattering experiments in high-energy particle accelerators had revealed a stunning proliferation of resultant particle-like excitations — 'resonances' — at particular energies, amounting to an ever-growing zoo of particles that could not possibly all be elementary (that is, indivisible).

It soon became apparent that the strongly interacting particles (also known as hadrons) could be ordered into certain symmetrical patterns akin to the periodic table of the elements. Together with evidence from scattering experiments that the most familiar hadrons — the protons and neutrons (nucleons) of the atomic nucleus — had a spatially extended structure, this finding led to the insight that hadrons are made from smaller particles, the quarks. But initial attempts to describe the forces between the quarks, and why they form the bound states they do, failed miserably.

So particle physicists started casting around for other ways of attacking the problem. In 1968, the Italian theoretician Gabriele Veneziano made a brilliant guess[9] and wrote down a concrete mathematical expression, the Veneziano amplitude, that explained some important features of high-energy scattering. But his formula could not be understood in terms of point-like particles; instead, it required the existence of extended objects — strings. These strings are thin tubes of energy formed by force lines that bind quarks together, and, just like violin strings, they can oscillate in many modes. The numerous resonances of strong-interaction physics would then be nothing but the different oscillation modes of these strings.

Unfortunately, this theory soon turned out to have several flaws, most seriously that, for mathematical consistency, a string must move in 25 spatial dimensions, rather than the familiar three. A rescue attempt, replacing the string with a new 'fermionic' variety with infinitesimal particle spins attached to the tubular-force lines, brought the 25 dimensions down to 9 — better, but still not good enough[10,11].

The arrival in the early 1970s of quantum chromodynamics (QCD), the quantum-field theory of the strong interaction, dealt the final blow to these early attempts to understand nuclear physics in terms of string theory. But, unfortunately, QCD is incredibly complex. Except for a few showpiece calculations — which put to rest any doubts that QCD might not be a correct description of the strong force — it is extremely difficult to extract measurable consequences from it. One of these calculations is the proof of 'asymptotic freedom', according

50 YEARS AGO

'Orbit of the artificial Earth satellite' — By a considerable feat of improvization, Mr. Martin Ryle and his team at the Mullard Radio Astronomy Observatory near Cambridge have been able to record the radio signals transmitted by the artificial satellite from October 5... A report from the United States gives the maximum height of the orbit above the Earth as 583 miles and the minimum as 143 miles, and states that the carrier rocket was travelling three minutes ahead of the satellite on October 12... Mr. D. H. Sadler reports... that the carrier rocket is now visible in the British Isles in the morning twilight... On October 13 it was approximately over Bournemouth at 5h. 26m. U.T., the track moving south-west, parallel to itself, about 200 km. a day.

From *Nature* 19 October 1957.

100 YEARS AGO

'Classification of portraits' — Experiments of various kinds that I have made to define the facial peculiarities of persons, families, and races by means of measurement led to the following results... The individuality of a portrait lies more in the relative positions of six cardinal features than in the shapes of the lines that connect them... The features are these:— *c*, the tip of the chin; *l*, the lower, and *u*, the upper lip; *m*, the hollow between the upper lip and the nose; *n*, the tip of the nose; *f*, the hollow between the nose and the brow... In my experiments I have chiefly used the side-view portraits by George Vance, R. A., of his distinguished contemporaries, published in 1809... I lexiconised these in respect to the measures...and found, first, that no two of the numerical formulae were the same... I have applied the above method to portraits of very different races, and have thus far found it efficient in all of them.

Francis Galton

From *Nature* 17 October 1907.

50 & 100 YEARS AGO

to which strong interactions become weak at very short distances. In this 'perturbative' regime, we understand (at least in principle) how to work with QCD. But for the strong coupling that occurs over larger distances, one has to resort to computer-simulation techniques, known as lattice QCD. These techniques have been rather successful (for instance, in explaining the spectrum of hadron masses), but rigorous results remain hard to come by: despite years of effort, we still cannot explain, for example, why there are no free, single quarks in nature. Such unresolved puzzles are coming into renewed focus with the scheduled start of experiments at the Large Hadron Collider at CERN in Geneva next year.

The new approach that revives the link to string theory first suggested itself in 1998, when Juan Martín Maldacena conjectured[12] a link between a close relative of QCD and a 'superstring' living in a ten-dimensional curved space-time. Although the theory in question, known as supersymmetric $N = 4$ gauge theory, is sufficiently different from QCD to be of no direct interest to experiment, the link raised the prospect of a general connection to some form of compactified string theory. This equivalence is now commonly referred to as the AdS/CFT (Anti-de-Sitter/conformal field theory) correspondence. If true, it would mean that string theory was originally not so far off the mark after all — its ingredients just need to be interpreted in the correct way.

The Maldacena conjecture raised a lot of interest, but seemed for a long time to be quantitatively unverifiable. This was because it takes the form of a duality in which the strongly coupled string theory corresponds to weakly coupled QCD-like theory, and vice versa. But to verify the duality, one would need to find a quantity to compare in a regime of intermediate coupling strength, and calculate it starting from both sides. No such quantity was obvious.

Help came from an entirely unexpected direction. Following a prescient observation[13], the spectrum of the $N = 4$ theory has been found[1,2] to be equivalently described by a quantum-mechanical spin chain of a type discovered by Hans Bethe in 1931 when modelling certain metallic systems. There are not many quantum-mechanical systems that can be solved analytically — the hydrogen atom is the most prominent example — but Bethe's *ansatz* immediately applied in a much wider context, and constructed a bridge between condensed-matter physics and string theory (in this context, see the recent News & Views article by Jan Zaanen[14] on the nascent connection to high-temperature superconductivity). Indeed, even though the mathematical description of the duality on the string-theory side is completely different from that on the condensed-matter side, a very similar, exactly solvable structure has been identified here as well[3–5].

Puzzling out the details of the exact solution is currently an active field of research. But in one instance, that idea had already been put to such a hard test that a complete solution now seems within reach. The context is a special observable entity, the 'cusp anomalous dimension', which was argued[6,7] to be ideally suited as a device to test whether string and gauge theory really connect. Some of its structure at strong coupling was also worked out. Just recently, Beisert, Eden and Staudacher[8] have extracted the analogue of this observable on the field-theory side, and have been able to write down an equation valid at any strength of the coupling. Since then, work has established that their 'BES equation' does indeed seem, for the first time, to offer a means of reformulating theories such as QCD as string theories.

Much still needs to be learned from this one exactly solvable case. There is justifiable hope that this solution will teach us how to go back to the physically relevant case of QCD and finally arrive at the long-sought dual description by a string theory. It may even take us closer to realizing the quantum-field theorist's ultimate dream, unfulfilled for more than 50 years: completely understanding an interacting relativistic quantum-field theory in the four space-time dimensions that we are familiar with. Progress towards this goal can be judged independently of loftier attempts to use strings in the construction of a theory of everything. ■

Hermann Nicolai is at the Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Mühlenberg 1, D-14476 Potsdam, Germany.
e-mail: nicolai@aei.mpg.de

1. Minahan, J. A. & Zarembo, K. *J. High Energy Phys.* **0303,** 013 (2003).
2. Beisert, N., Kristjansen, C. & Staudacher, M. *Nucl. Phys. B* **664,** 131–184 (2003).
3. Bena, I., Polchinski, J. & Roiban, R. *Phys. Rev. D* **69,** 046002 (2004).
4. Kazakov, V. A., Marshakov, A., Minahan, J. A. & Zarembo, K. *J. High Energy Phys.* **0405,** 024 (2004).
5. Arutyunov, G., Frolov, S. & Staudacher, M. *J. High Energy Phys.* **0410,** 016 (2004).
6. Gubser, S. S., Klebanov, I. R. & Polyakov, A. M. *Nucl. Phys. B* **636,** 99–114 (2002).
7. Frolov, S. & Tseytlin, A. A. *J. High Energy Phys.* **0206,** 007 (2002).
8. Beisert, N., Eden, B. & Staudacher, M. *J. Stat. Mech.* P01021 (2007).
9. Veneziano, G. *Nuovo Cimento* **57A,** 190 (1968).
10. Ramond, P. *Phys. Rev. D* **3,** 2415–2418 (1971).
11. Neveu, A. & Schwarz, J. H. *Nucl. Phys. B* **31,** 86–112 (1971).
12. Maldacena, J. M. *Adv. Theor. Math. Phys.* **2,** 231–252 (1998).
13. Lipatov, L. N. preprint available at www.arxiv.org/abs/hep-th/9311037 (1993).
14. Zaanen, J. *Nature* **448,** 1000–1001 (2007).

MICROBIOLOGY

Preparing the shot

Christof R. Hauck

Direct injection of proteins into host cells is one of the tricks bacteria use during infection. It seems that, to achieve this, the stomach pathogen *Helicobacter pylori* first grabs the cell by its surface receptors.

The bacterium *Helicobacter pylori* successfully colonizes the stomach of about every third person. Infection with this ubiquitous microorganism can cause acute and chronic gastritis, as well as stomach ulcers[1]. Moreover, up to 90% of cases of stomach cancer are associated with *H. pylori* infection. The bacterium's main weapon is an elaborate apparatus on its surface called the type-IV secretion system, which acts as a nano-syringe (Fig. 1a). Using this apparatus, the bacterium delivers a cancer-associated protein, CagA, directly into its host cells. But whether the bacterium anchors the secretion system to the surface of host cells before injection, and if so, how, has remained unclear. On page 862 of this issue, Kwok *et al.*[2] report that transfer of CagA is made possible by another *H. pylori* protein, CagL, which binds to integrin receptors on gastric epithelial cells.

So far, CagA is the only *H. pylori* protein known to be injected into the host cell. In the bacterial chromosome, the *cagA* gene is part of a stretch of DNA called *cag*PAI, which also encodes the structural components of the type-IV secretion machinery[3]. Bacterial strains harbouring *cag*PAI are considered to be more virulent than other strains[4].

Previous work[5–7] had shown that, once CagA is delivered into the host cell, kinase enzymes of the Src family add a phosphate group to it. The presence of phosphorylated CagA results in several changes that might promote *H. pylori* virulence and an unfavourable outcome for infection with this bacterium[4,8]. These changes include the assembly of signalling complexes, reduced cell–cell adhesion and induction of cell migration.

Examining the localization of phosphorylated CagA in isolated gastric epithelial cells, Kwok *et al.*[2] found that it occurs almost exclusively at focal adhesion sites — discrete regions of the cell where integrin receptors 'glue' cells to the supporting extracellular matrix. The authors speculated that CagA might not move through the cytoplasm of the infected cells to these sites, but instead be injected directly at these places. Support for this idea came from experiments demonstrating that CagA is not transferred into host cells if *H. pylori* cannot

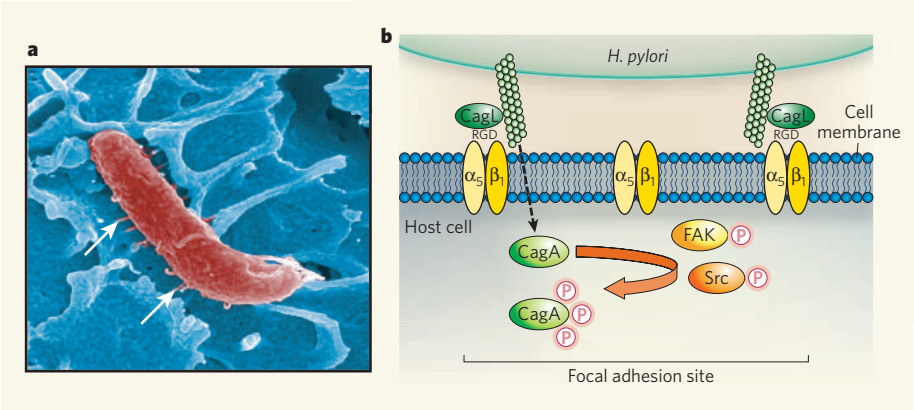


Figure 1 | Translocation of the CagA protein of the bacterium *Helicobacter pylori* into host cells. **a**, Electron micrograph of *H. pylori* (red) attaching to the surface of a host gastric epithelial cell. The type-IV secretion apparatus, in the form of needle-like projections (arrowed), mediates the transfer of the bacterial protein CagA into the host cell. (Image courtesy of M. Rohde.) **b**, Kwok *et al.*[2] find that, to achieve transfer, the CagL protein on the surface of the secretion apparatus binds to host-cell integrins ($\alpha_5\beta_1$) through its RGD structural motif, thereby directing the injection of CagA precisely into focal adhesion sites in the cell. Moreover, CagL seems to stimulate integrin-mediated signalling, such as activation of FAK and Src-family kinases, ensuring the phosphorylation of translocated CagA, and thus its activation.

access a particular type of integrin known as integrin $\alpha_5\beta_1$.

The authors went on to show that CagL — a little-studied protein encoded by a sequence within *cag*PAI — is the crucial molecular link between the type-IV secretion apparatus of *H. pylori* and host-cell integrins, mediating CagA translocation. CagL occurs on the surface of *H. pylori* together with the type-IV secretion apparatus. It contains a three-amino-acid sequence motif known to mediate binding to, and activation of, integrin $\alpha_5\beta_1$. This tripeptide motif in CagL consists of arginine-glycine-aspartate (RGD in single-letter code) — a sequence that is present in several proteins found in the extracellular matrix, such as fibronectin and vitronectin, that bind to integrins to activate them[9].

Kwok and colleagues showed that, like the physiological stimulation of integrins, normal, RGD-containing CagL — but not mutant CagL (containing RGA instead) — activates integrin-associated kinases such as FAK and Src, which then phosphorylate CagA (Fig. 1b). Therefore, CagL not only holds on to integrins to precisely position the type-IV secretion apparatus, but, by stimulating integrin-mediated signalling, it also ensures that the translocated CagA encounters active host kinases for its subsequent phosphorylation.

A number of pathogenic bacteria enter host cells by exploiting integrins and integrin-initiated signalling events[10]. So a remaining question is how *H. pylori* avoids being internalized in this way. In the intact tissue, preformed integrin-rich focal adhesion sites occur on the basolateral surface of gastric epithelial cells, facing away from the stomach lumen. Other unknowns are therefore how, once in the stomach, *H. pylori* reaches integrin-rich focal adhesion sites on the basolateral surface, and how CagL induces the *de novo* formation of integrin clusters on the cell's apical surface, which faces the stomach lumen.

A type-IV secretion apparatus or a functionally similar type-III secretion system is present in many other bacteria, and interactions with host-cell integrins have been documented[11]. So the race will now be on to search for CagL-related proteins in these microbes. Furthermore, to understand how CagL helps position the *H. pylori* type-IV secretion apparatus, it will be crucial to dissect the interactions between CagL and the structural components of this apparatus.

Regardless of the unresolved questions, Kwok and colleagues' work[2] neatly highlights the thrifty elegance with which a bacterial pathogen uses a single virulence factor, together with exploitation of the host-cell components, to secure intracellular positioning and activation of the noxious CagA protein. ■

Christof R. Hauck is at the Lehrstuhl Zellbiologie, Fachbereich Biologie X908, Universität Konstanz, Universitätsstrasse 10, 78457 Konstanz, Germany.
e-mail: christof.hauck@uni-konstanz.de

1. Kusters, J. G., van Vliet, A. H. & Kuipers, E. J. *Clin. Microbiol. Rev.* **19,** 449–490 (2006).
2. Kwok, T. *et al. Nature* **449,** 862–866 (2007).
3. Censini, S., Stein, M. & Covacci, A. *Curr. Opin. Microbiol.* **4,** 41–46 (2001).
4. Hatakeyama, M. *Nature Rev. Cancer* **4,** 688–694 (2004).
5. Stein, M., Rappuoli, R. & Covacci, A. *Proc. Natl Acad. Sci. USA* **97,** 1263–1268 (2000).
6. Odenbreit, S. *et al. Science* **287,** 1497–1500 (2000).
7. Selbach, M., Moese, S., Hauck, C. R., Meyer, T. F. & Backert, S. *J. Biol. Chem.* **277,** 6775–6778 (2002).
8. Amieva, M. R. *et al. Science* **300,** 1430–1434 (2003).
9. Xiong, J. P. *et al. Science* **296,** 151–155 (2002).
10. Agerer, F. *et al. J. Cell Sci.* **118,** 2189–2200 (2005).
11. Watarai, M., Funato, S. & Sasakawa, C. *J. Exp. Med.* **183,** 991–999 (1996).

ASTRONOMY

Black holes go extragalactic

Tomasz Bulik

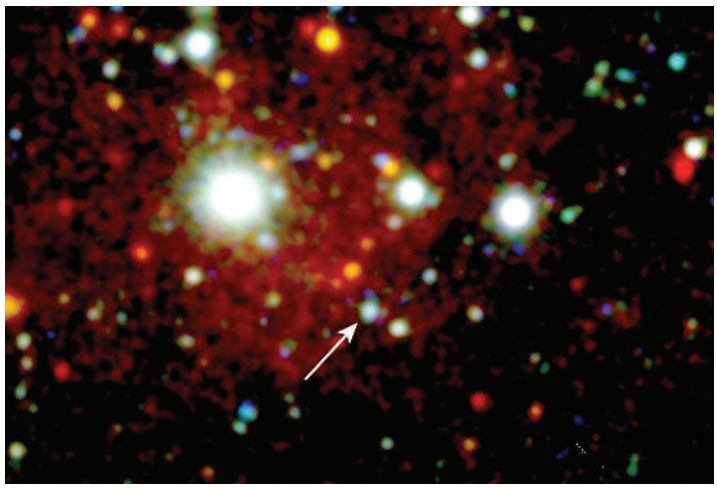
The mass of a black hole beyond our Galaxy has been calculated, thanks to the presence of an associated star. The hole is the weightiest yet, placing intriguing constraints on how this binary system developed.

Weighing celestial bodies is a goal of many astronomers, and it is an especially fascinating task in the case of black holes. On page 872 of this issue[1], Orosz *et al.* announce a significant advance — they have measured the mass of a black hole lying far beyond our own Galaxy.

The best method for weighing a black hole is to measure the strength of its gravity. This can be done by observing how it affects the motion of a second object, usually a blob of hot gas, or a star, that is locked into a binary system with it. It's a well-worn method: the radial velocity and orbital period of the two objects must first be measured, along with their eccentricity (the deviation of their orbits from a perfect circle). This measurement sets a joint constraint on the masses of the objects and their inclination (the angle between their orbital paths and some reference plane). But it does not tell us what the values of the masses and inclination are.

In the case of eclipsing binaries, where one object periodically passes in front of the other, the inclination can be tightly constrained: in a binary, both objects must, by definition, be observed nearly in the plane of their orbits. But such instances are rare, and the lack of reliable information on the orbital inclination of most black holes remains a great source of uncertainty in the measurement of their mass. A final ingredient for the calculation is a direct measurement of the mass of the companion object. If it is a star, an estimate can be made from its luminosity and the spectrum of radiation that it emits. Astronomers cannot measure the absolute luminosity directly, only its radiation flux at Earth; so they first need an estimate of its distance from us.

Several binary systems hosting extremely compact, massive objects are known in our Galaxy, as revealed by their characteristic high and variable luminosity at X-ray wavelengths. None is eclipsing, but the objects' masses have been determined using the classic method outlined above. The values cluster at around 10 times the mass of the Sun, significantly higher than the 2.5 to 3 solar masses thought to



W. PIETSCH, MPE/ESA

Figure 1 | In a galaxy quite close by. The M 33 X-7 binary system (arrowed) in which Orosz and colleagues[1] weigh the black hole is part of the M 33, or Triangulum, galaxy, a member of the Milky Way's Local Group.

be the maximum for a neutron star, the dense core of a burnt-out massive star. Thus, according to our models of such things, the compact objects must be black holes.

Orosz *et al.*[1] investigate in detail a similar X-ray binary, known as M 33 X-7, in a prominent member of the Local Group of galaxies, called Triangulum or M 33 (Fig. 1). The source was discovered[2] by NASA's Einstein satellite in 1981, and suggested to be an eclipsing binary[3] eight years later. It has since been observed by many satellites, and the suspicion that it is hosting a black hole has grown all the time. Orosz and colleagues identify the companion star in the binary and measure its orbital motion, thus constraining its mass. Because this is an eclipsing binary, the inclination of the second object's orbit is known precisely, and so its mass can readily be determined. The result is 15.65 solar masses — making it not just a black hole, but the most massive black hole of stellar origin measured so far. (Other, 'supermassive' black holes that are thought to lurk at the centre of most, if not all, galaxies have masses of millions to billions of Suns, but are quite different beasts.)

The uncertainty on this mass is unusually small — just 1.45 solar masses. This is not just because the inclination of the black hole is well constrained, but also, counter-intuitively, because it is so far away. That means that its distance from us can be determined much more precisely than for binaries in our own Galaxy: once an object has been identified to belong to a particular galaxy, the uncertainty in its distance is due only to the ratio of the galaxy's radial distance from Earth to its absolute distance from Earth. This is less than a few per cent, whereas for binaries in our Galaxy the same uncertainty can easily reach a factor of two. Such greater accuracy feeds through the calculation of the companion star's luminosity and mass, and can thus bring a more precise determination of the black hole's mass than for binaries in the Milky Way.

So where does all this leave us? The range of black-hole masses is still poorly documented, but Orosz and colleagues' findings[1] mean that it is likely to extend to higher values than currently known. By making additional, more precise measurements of the masses of black holes in binaries, and extending the surveys to other galaxies, we will gain a better understanding of how black holes form. This process is thought to occur in the gravitational collapse of stars of more than some 20 solar masses. But how big is the most massive black-hole mass that can form in such a collapse? And does it depend on the star's metallicity — its relative abundance of elements heavier than hydrogen and helium?

The small size of the M 33 X-7 system indicates that it has gone through a violent stage of evolution called the common-envelope phase, in which one object sucks the other inside and at the same time expels its own outer layers (Fig. 2). This leads either to a merger of the two bodies or the formation of a tight binary in which one star is stripped of its outer layers. Although common among binaries, this stage is extremely difficult to observe directly, as it lasts only a few hundred years. During the common-envelope phase, the progenitor of the M 33 X-7 black hole must have lost a large amount of mass for the two objects to be so close; but on the other hand, it must have retained enough mass to form such a heavy black hole. This system might thus provide both the upper and lower limits on the amount of mass loss and orbital tightening that can occur in the common envelope.

Such constraints are important in the continuing search for the ripples in space-time known as gravitational waves, as they will determine the number and properties of binaries consisting of two black holes that have survived the common-envelope phase. These systems would be the strongest sources of gravitational waves in the cosmos. But too much orbital tightening during the common-envelope phase leads to a fast merger of the two objects, whereas too little would lead to the formation of a widely separated system with no chance of merging. Two gravitational-wave observatories, LIGO in the United States and VIRGO in Italy, are currently primed to detect these ripples in space-time, but have yet to make any positive detection.

Extending the study of compact-object binaries to the extragalactic population is a giant leap forwards, both in terms of the number of objects that can be studied and the accuracy with which we can characterize them. Orosz and colleagues' discovery[1] is just the beginning. ■

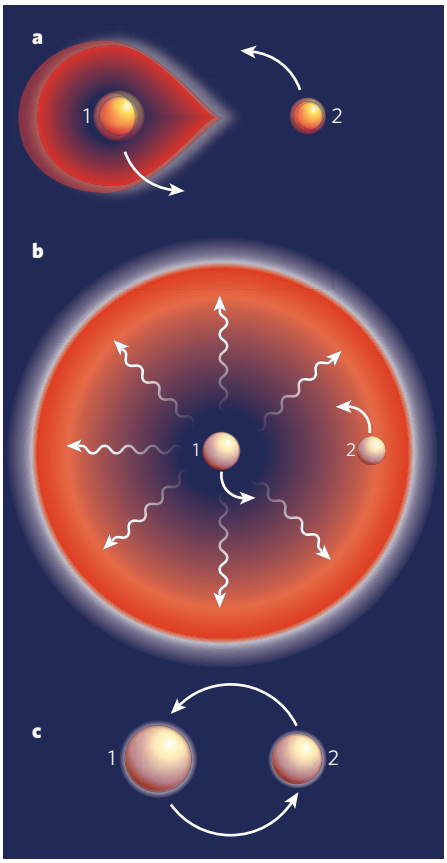
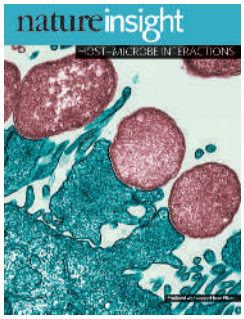


Figure 2 | Stuffing a common envelope. The black hole and the orbiting star that make up the M 33 X-7 binary system are so close to each other that they must have gone through a phase of evolution known as the common envelope. **a**, First, the more massive of two stars (star 1) expands into its red-giant phase towards the end of its life, entirely filling its 'Roche lobe' — the volume within which the gravitational attraction of the star is great enough for any released material to fall back onto its surface. **b**, This expansion results in the establishment of a common envelope of gravitational attraction around the two stars, and sends the smaller star spiralling in towards the first. But frictional forces between the common envelope and the star cause the enveloping material of star 1 to be expelled. **c**, Eventually, the envelope disappears, leaving a tight system consisting of the core of star 1 in a tight orbit with star 2. Later, the spent core of star 1 collapses to form a black hole, completing the system's evolution so far.

Tomasz Bulik is at the Astronomical Observatory, University of Warsaw, Al Ujazdowskie 4, PL-00-478 Warsaw, and the Copernicus Astronomical Center, Warsaw, Poland.
e-mail: tb@astrouw.edu.pl

1. Orosz, J. A. *et al. Nature* **449,** 872–875 (2007).
2. Long, K. S., Dodorico, S., Charles, P. A. & Dopita, M. A. *Astrophys. J.* **246,** L61–L64 (1981).
3. Peres, G., Reale, F., Collura, A. & Fabbiano, G. *Astrophys. J.* **336,** 140–151 (1989).



HOST–MICROBE INTERACTIONS

Cover illustration
A human epithelial cell infected with enteropathogenic *Escherichia coli*. (Courtesy of J. A. Guttman, A. W. Vogl and B. B. Finlay.)

Editor, *Nature*
Philip Campbell
Insights Publisher
Sarah Greaves
Publishing Assistant
Claudia Banks
Insights Editor
Ritu Dhand
Production Editor
Davina Dadley-Moore
Senior Art Editor
Martin Harrison
Art Editor
Nik Spencer
Sponsorship
Emma Green
Production
Susan Gray
Marketing
Katy Dunningham
Elena Woodstock
Editorial Assistant
Alison McGill

More than a century ago, Robert Koch established that infectious diseases are caused by microbes, a discovery that won him the Nobel Prize in Physiology or Medicine in 1905. At around the same time, Ilya Mechnikov, one of the pioneers of cellular immunology, was the first to recognize that microbes might also have beneficial effects on human health, when he proposed that 'lactic-acid bacteria' can prolong life.

Since then, a tremendous amount has been discovered about encounters between microbes and the animals they colonize — their hosts. Host–microbe interactions are as diverse as the organisms involved: they can be accidental or obligatory; they can result in temporary or persistent intimate associations; and they can involve subtle or intense molecular and cellular responses. But the outcome for the host is simple: health or disease.

In the quest to understand and combat infectious diseases and, more recently, to uncover the basis of non-pathogenic microbial colonization, microbes have been found to produce a multitude of factors that either confer virulence or promote colonization by other means. The actions of these factors are countered by the equally diverse responses of the host immune system. This Insight highlights advances in the study of this dynamic interplay between host and microbe, focusing on humans and bacteria. It also provides an overview of the current understanding of the ecology, evolution, immunology, cell biology and genomics of these interactions. We thank the authors and reviewers, who contributed their time, effort and enthusiasm to this collection.

We are pleased to acknowledge the financial support of Pfizer in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

Claudia Lupp, Senior Editor

FEATURE

REVIEWS

nature insight

NATURE|Vol 449|18 October 2007|doi:10.1038/nature06244

The Human Microbiome Project

Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight & Jeffrey I. Gordon

A strategy to understand the microbial components of the human genetic and metabolic landscape and how they contribute to normal physiology and predisposition to disease.

Before the Human Genome Project was completed, some researchers predicted that ~100,000 genes would be found. So, many were surprised and perhaps humbled by the announcement that the human genome contains only ~20,000 protein-coding genes, not much different from the fruitfly genome. However, if the view of what constitutes a human is extended, then it is clear that 100,000 genes is probably an underestimate. The microorganisms that live inside and on humans (known as the microbiota) are estimated to outnumber human somatic and germ cells by a factor of ten. Together, the genomes of these microbial symbionts (collectively defined as the microbiome) provide traits that humans did not need to evolve on their own[1]. If humans are thought of as a composite of microbial and human cells, the human genetic landscape as an aggregate of the genes in the human genome and the microbiome, and human metabolic features as a blend of human and microbial traits, then the picture that emerges is one of a human 'supraorganism'.

To understand the range of human genetic and physiological diversity, the microbiome and the factors that influence the distribution and evolution of the constituent microorganisms must be characterized. This is one of the main goals of the Human Microbiome Project (HMP). The outcome might also provide perspective on contemporary human evolution: that is, on whether and how rapidly advancing technology, and the resultant transformation of human lifestyles and the biosphere, influences the 'micro-evolution' of humans and thereby health and predisposition to various diseases.

The HMP is a logical conceptual and experimental extension of the Human Genome Project. The HMP is not a single project. It is an interdisciplinary effort consisting of multiple projects, which are now being launched concurrently worldwide, including in the United States (as part of the next phase of the National Institutes of Health's Roadmap for Medical Research), Europe and Asia. The advent of highly parallel DNA sequencers and high-throughput mass spectrometers with remarkable mass accuracy and sensitivity is propelling microbiology into a new era, extending its focus from the properties of single organism types in isolation to the operations of whole communities. The new field of metagenomics involves the characterization of the genomes in these communities, as well as their corresponding messenger RNA, protein and metabolic products[2].

The HMP will address some of the most inspiring, vexing and fundamental scientific questions today. Importantly, it also has the potential to break down the artificial barriers between medical microbiology and environmental microbiology. It is hoped that the HMP will not only identify new ways to determine health and predisposition to diseases but also define the parameters needed to design, implement and monitor strategies for intentionally manipulating the human microbiota, to optimize its performance in the context of an individual's physiology. Examples of, and speculations about, the functional contributions of the microbiota are provided in Box 1.

In this article, we discuss the conceptual and experimental challenges that the HMP faces, as well as the rewards it might hold. We focus on the gut when providing examples, because this habitat harbours the largest collection of microorganisms.

Ecology and considerations of scale

Questions about the human microbiome are new only in terms of the system to which they apply. Similar questions have inspired and confounded ecologists working on macroscale ecosystems for decades. It is expected that the HMP will uncover whether the principles of ecology, gleaned from studies of the macroscopic world, apply to the microscopic world that humans harbour (see page 811). In particular, the following questions might be answered by the HMP. How stable and resilient is an individual's microbiota throughout one day and during his or her lifespan? How similar are the microbiomes between members of a family or members of a community, or across communities in different environments? Do all humans have an identifiable 'core' microbiome, and if so, how is it acquired and transmitted? What affects the genetic diversity of the microbiome (Fig. 1), and how does this diversity affect adaptation by the microorganisms and the host to markedly different lifestyles and to various physiological or pathophysiological states?

To address any question about the human microbiome, the microbiota needs to be sampled, and temporal and spatial scales need to be considered before undertaking this process. For example, microbial communities on human surfaces (that is, the skin and mucosal surfaces such as the gut) have a complex biogeography that can be defined at a range of distances: at the micrometre scale (the distribution of microorganisms on undigested food particles in the distal gut or across a mucosal barrier); at the centimetre scale (the distribution of communities around different teeth); and at the metre scale (the distribution of communities along the long axis of the gut).

Scale also has a further meaning. The core microbiome is whatever factors are common to the microbiomes of all or the vast majority of humans. At present, there are 6.7 billion humans on Earth. Because of various constraints, the human microbiome(s) will need to be characterized by comparing limited data types collected from a limited set of individuals. If human body habitats, such as the gut, are viewed as 'islands' in space and time, then island-biogeography theory, which was developed from studies of macroscale ecosystems[3], might be useful for understanding the observed microbial diversity. This theory states that community composition can depend strongly on the order in which species initially enter a community (a phenomenon known as multiple stable states[4]). The importance of the initial inoculating microbial community on the community composition at later stages is evident from animal studies. For example, in the mouse gut microbiota, the effects of maternal transmission (kinship) are apparent over several generations in animals of the same inbred strain[5]. Similarly, experiments in which the microbiota is transferred from one host to another, from conventionally raised mice or zebrafish to germ-free mice or zebrafish, demonstrate that the microbial community available to colonize the gut at the time of birth, together with the features of the gut habitat itself, conspire to select a microbiota[6]. To study the human microbiome, a few specific islands (humans) could be characterized in depth. Alternatively, the equivalent of a biogeography experiment could be carried out, in which general trends are inferred from a coarse-grained analysis of a larger number of humans, who are selected on the basis of demographic, geographical or epidemiological

factors. These strategies are complementary and, as discussed later, both will be needed to understand the human microbiome fully.

What do we know about the human microbiome?

Although the human microbiome is largely unexplored, recent studies have begun to reveal some tantalizing clues about its features.

Large variation in bacterial lineages between people

The decreasing cost and increasing speed of DNA sequencing, coupled with advances in the computational approaches used to analyse complex data sets[7–11], have prompted several research groups to embark on small-subunit (16S) ribosomal RNA gene-sequence-based surveys of bacterial communities that reside on or in the human body, including on the skin and in the mouth, oesophagus, stomach, colon and vagina[12–17] (see page 811). The 16S rRNA gene is found in all microorganisms and has enough sequence conservation for accurate alignment and enough variation for phylogenetic analyses. The largest reported data sets are for the gut, although the number of people sampled by using these culture-independent surveys is still limited. Most of the 10–100 trillion microorganisms in the human gastrointestinal tract live in the colon. More than 90% of all phylogenetic types (phylotypes) of colonic bacteria belong to just 2 of the 70 known divisions (phyla) in the domain Bacteria: the Firmicutes and the Bacteroidetes. For samples taken from the colon, the differences between individuals are greater than the differences between different sampling sites in one individual[15]. Moreover, faeces are representative of interindividual differences[5]. A recent study of 18,348 faecal 16S rRNA gene sequences collected from 14 unrelated adults over the course of a year showed large differences in microbial-community structure between individuals, and it established that community membership in each host was generally stable during this period[16]. How is such high interindividual diversity sustained? The observations about diversity in the human gut microbiota might fit with predictions of the neutral theory of community assembly, which states that most species share the same general niche (an ecological term that, in the case of microorganisms, refers to 'profession'), or the biggest niche, and therefore are likely to be functionally redundant[18]. Therefore, this theory predicts that highly variable communities (as defined by 16S rRNA gene lineages) will have high levels of functional redundancy between community members.

Ecosystem-level functions

Comparative metagenomics has uncovered functional attributes of the microbiome. The first reported application of metagenomic techniques to a human microbiome involved two unrelated, healthy adults. Compared with all previously sequenced microbial genomes and the human genome, metabolic reconstructions of the gut (faecal) microbiomes of these adults showed significant enrichment for genes involved in several metabolic pathways: the metabolism of xenobiotics (that is, foreign substances), glycans and amino acids; the production of methane; and the biosynthesis of vitamins and isoprenoids through the 2-methyl-D-erythritol 4-phosphate pathway[1].

The usefulness of comparative metagenomics is further underscored by a recent study, which showed that a host phenotype (obesity) can be correlated with the degree of representation of microbial genes involved in certain metabolic pathways[19]. Microbial-community DNA was isolated from the distal-gut contents of genetically obese animals (*ob/ob* mice, which have a mutation in the gene encoding leptin) and their lean littermates (+/+ or *ob*/+) and then sequenced. Predictions of microbial-community metabolism, based on community gene content, indicated that the obesity-associated gut microbiome has an increased capacity to harvest energy from the diet. Specifically, the *ob/ob* mouse microbiome was enriched for genes involved in importing and metabolizing otherwise indigestible dietary polysaccharides to short-chain fatty acids, which are absorbed by the host and stored as more complex lipids in adipose tissue. Biochemical analyses supported these predictions. Moreover, when adult germ-free wild-type mice were colonized with a gut microbiota from obese (*ob/ob*) or lean (+/+) mice, adiposity

Box 1 | Examples of functional contributions of the gut microbiota

Harvest of otherwise inaccessible nutrients and/or sources of energy from the diet, and synthesis of vitamins

The nutrient and/or energetic value of food is not absolute but is affected, in part, by the digestive capacity of an individual's microbiota[1,19,42–44]. This has implications for identifying individuals who are at risk of being malnourished or obese and treating them on the basis of a more personalized view of nutrition that considers their microbial ecology.

Metabolism of xenobiotics, and other metabolic phenotypes

The microbiota is a largely underexplored regulator of drug metabolism and bioavailability. Bioremediation-like functions of the microbiota, such as detoxifying ingested carcinogens, might affect a host's susceptibility to various neoplasms, both within and outside the gut. In addition, the metabolism of oxalate by the microbiota has been linked to a predisposition to the development of kidney stones[45]. Also, the modification of bile acids by microorganisms affects lipid metabolism in the host[44]. Ascribing metabolic phenotypes (also known as metabotypes) to the microbiota should extend our repertoire of personalized biomarkers of health and of disease susceptibility.

Renewal of gut epithelial cells

The renewal of gut epithelial cells is affected, in part, by interactions between the microbiota and immune cells. Effects could range from susceptibility to neoplasia[46] to the capacity for repairing a damaged mucosal barrier[47]. Germ-free mice renew gut epithelial cells at a slower rate than their colonized counterparts[47]. Comparing microbial communities that are physically associated with neoplasms and those with varying degrees of remoteness from the neoplasms might provide new mechanistic insights about cancer pathogenesis.

Development and activity of the immune system

The gut microbial community has an effect on both the innate immune system[48] and the adaptive immune system[49], and it contributes to immune disorders that are evident within and outside the gut. For example, in individuals with inflammatory bowel diseases, the immune response to the gut microbial community seems to be dysregulated: genome-wide association studies of patients with Crohn's disease have identified several human genes involved in both innate and adaptive immune responses[50]. In addition, susceptibility to colonization by enteropathogens is affected by the capacity of the microbiota to alter the expression of host genes encoding antimicrobial compounds[48,51]. Furthermore, the incidence of asthma is correlated with exposure to bacteria during childhood[52] and treatment with broad-spectrum antibiotics in early childhood[53].

Cardiac size

Germ-free animals have a smaller heart as a proportion of body weight than their colonized counterparts[54]. The mechanism underlying this phenotype has yet to be defined, but this finding emphasizes the importance of studying the extent to which human physiology is modulated by the microbiome.

Behaviour

Germ-free mice have greater locomotor activity than their colonized counterparts[43]. It will be interesting to study whether there are behavioural effects in humans. Has the microbiota evolved ways to benefit itself and its host by influencing human behaviour? Is altered production of neurologically active compounds (either directly, by the microbiota, or indirectly, by microbiota-mediated modulation of the expression of host genes that encode products normally involved in the biosynthesis and/or metabolism of these compounds) associated with any neurodevelopmental and/or psychiatric disorders?

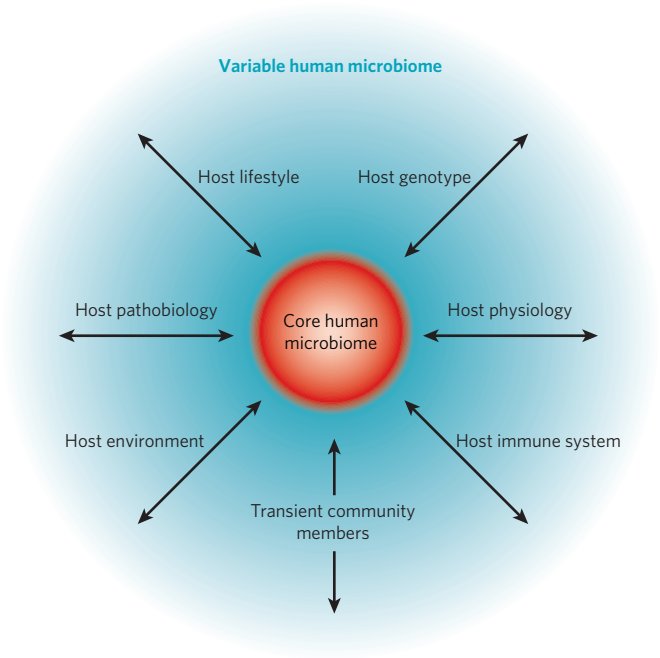


Figure 1 | The concept of a core human microbiome. The core human microbiome (red) is the set of genes present in a given habitat in all or the vast majority of humans. Habitat can be defined over a range of scales, from the entire body to a specific surface area, such as the gut or a region within the gut. The variable human microbiome (blue) is the set of genes present in a given habitat in a smaller subset of humans. This variation could result from a combination of factors such as host genotype, host physiological status (including the properties of the innate and adaptive immune systems), host pathobiology (disease status), host lifestyle (including diet), host environment (at home and/or work) and the presence of transient populations of microorganisms that cannot persistently colonize a habitat. The gradation in colour of the core indicates the possibility that, during human micro-evolution, new genes might be included in the core microbiome, whereas other genes might be excluded.

increased to a significantly greater degree in recipients of the microbiota from obese mice than in recipients of the microbiota from lean mice, supporting the conclusion that the obesity-associated gut microbiota has an increased (and transmissible) capacity to promote fat deposition[19]. This coupling of comparative metagenomics with germ-free animal models shows one way to proceed from *in silico* predictions to experimental tests of whole-community microbiome function.

Metagenomic data sets from different microbial ecosystems can also be compared, allowing the traits that are important to each to be uncovered[20]. An example of such an analysis is shown in Fig. 2. The human and mouse gut-microbiome data sets described in this section are compared with data sets obtained from three environmental communities: decaying whale carcasses located at the bottom of the ocean (known as whale falls), an agricultural-soil community and a survey of the Sargasso Sea[20,21]. DNA-sequencing reads were culled from each data set and matched to annotated genes in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[22]. The gut microbiomes were found to cluster together and, compared with the environmental microbiomes, are enriched for predicted genes assigned to KEGG categories and pathways for carbohydrate and glycan metabolism (Fig. 2). Deeper sequencing of more human gut microbiomes will be required to determine whether these features are common traits of the human microbiome. (For further discussion of sampling issues, see the section Designing comparisons of microbial communities in humans.)

What will the HMP need for success?

Several factors will need to come together as this international effort is launched.

Sequencing more reference genomes

At present, metagenomic analyses of complex microbial communities are limited by the availability of suitable reference genomes, which are needed for confident assignment of the short sequences produced by the current generation of highly parallel DNA sequencers. These analyses are also constrained by a lack of knowledge about the niches of the organismal lineages that constitute these communities. An ongoing project to sequence the genomes of 100 cultured representatives of the phylogenetic diversity in the human gut microbiota[23] illustrates how reference genomes will help to interpret metagenomic studies. Capillary-sequencing reads from the human and mouse gut-microbiome data sets described earlier were matched to published microbial and eukaryotic genomes (KEGG database version 40 (ref. 22)) and 17 recently sequenced genomes of human gut bacteria (http://genome.wustl.edu/pub/) belonging to the divisions Bacteroidetes, Firmicutes and Actinobacteria (BLASTX best-BLAST-hit E value $< 10^{-5}$; http://www.ncbi.nlm.nih.gov/BLAST). These analyses showed that the quality of the sequence matches and the proportion of metagenomic read assignments increases with the inclusion of each additional gut bacterial genome.

The sequencing of more reference genomes, including genomes from multiple isolates of selected species-level phylotypes, should also help to answer questions about genetic variation within and between the major phylogenetic lineages in a given habitat, such as the gut. For example, a comparison of members of the Firmicutes and Bacteroidetes should provide insight into the extent of genetic redundancy and/or specialization between these two divisions. Given the extraordinary density of colonization in the distal gut (10^{11}–10^{12} organisms per ml of luminal content), these extra genomes would also provide an opportunity to determine more accurately the role of horizontal gene transfer in the evolution of gut microorganisms within and between hosts[24], as well as the extent to which the gene content of these microorganisms reflects their phylogenetic history.

To obtain reference-genome sequences, it will be crucial to develop new methods for retrieving microorganisms that cannot be cultured at present. Recently, several methods — fluorescence *in situ* hybridization with phylogenetic markers, flow cytometry, and whole-genome amplification and shotgun sequencing — have been used to obtain a partial genome assembly for a member of the candidate phylum TM7, providing a first look at a group of microorganisms with no culturable representatives[25]. In addition, methods such as the encapsulation of cells in gel microdroplets are aimed at enabling high-throughput culture of microorganisms in a simulated natural environment[26].

Linking short gene fragments to organisms

Because metagenomic data sets consist largely of unassembled sequence data, another major challenge is to link genes to organisms or at least to broader taxonomic classifications. Several approaches exist[27–29], but no tools have been developed for the automated analysis of large data sets containing mostly short sequence reads, without relying on phylogenetic marker genes. Thus, developing an accurate and scalable way to phylogenetically classify huge numbers of short sequence reads is essential.

The two general marker-independent approaches to phylogenetic assignment are to use Markov models based on the frequency of short nucleotide sequences (or 'words') in the reads and to use homology searching to place each sequence fragment in the context of a phylogenetic tree. Because of statistical sampling issues, the Markov-model-based approach is likely to be relatively insensitive, especially for short sequences and for sequences from heterogeneous genomes. The homology-search-based approach is probably more accurate and provides the additional advantage of placing each sequence in the context of a multiple alignment and a phylogenetic tree, which can then be used in further studies. However, sequences without identifiable homologues cannot be analysed in this way. A combination of these two general strategies is likely to be the best approach to understanding the functions associated with each metagenome.

There are three key issues when considering these approaches. First, it is important to understand how accurate the phylogenetic classification obtained by using each method can be, especially in the face of horizontal gene transfer. Second, it will be necessary to find better, faster and more scalable heuristics for generating huge phylogenetic trees that contain millions of sequences. Third, it is important to identify the best way to account for the effects of both the genome and the function of each encoded protein on the overall composition of each sequence. In particular, heterogeneous rates of evolution in different protein families pose substantial problems for search-based methods: considerable similarities at the primary-structure level might not persist over time, and the secondary and tertiary structures of the proteins are usually unknown, thus preventing the use of structure-based alignment techniques.

Designing comparisons of microbial communities in humans

Understandably, there will be great pressure at early stages of the HMP to focus on disease states. However, 'normal' states need to be defined before the effect of the microbiota on disease predisposition and pathogenesis can be evaluated, and this will require time, resources and discipline.

Several issues need to be considered when designing ways to generate an initial set of reference microbiomes from healthy individuals. What is the degree of genetic relatedness between those who are sampled: for example, should the initial focus be on monozygotic and dizygotic twins and their mothers? What is the place of the sampled individuals in the family structure? What age are they, and what are their demographics (for example, rural versus urban environment and lifestyle)? What are the ethical, legal and logistical barriers that need to be overcome to obtain, without exploitation, samples and metadata (that is, 'relevant' environmental and host parameters) from people with diverse cultural and socio-economic backgrounds? What types of comparison are needed: for example, should there be measurements of diversity within samples (α diversity); between samples (β diversity); between body habitats in a given individual; and/or between family members for a given habitat? And what protocols could or should be used for sampling surface-associated microbial communities? This last issue is a major unresolved technical problem. At present, there are no methods to retrieve sufficient quantities of microorganisms from various body surfaces, such as the skin and the vaginal mucosa, in a reproducible and representative manner, and sufficiently free of human cells, so that the microbiome can be sequenced. It is also unclear at what temporal and spatial scales this sampling should occur.

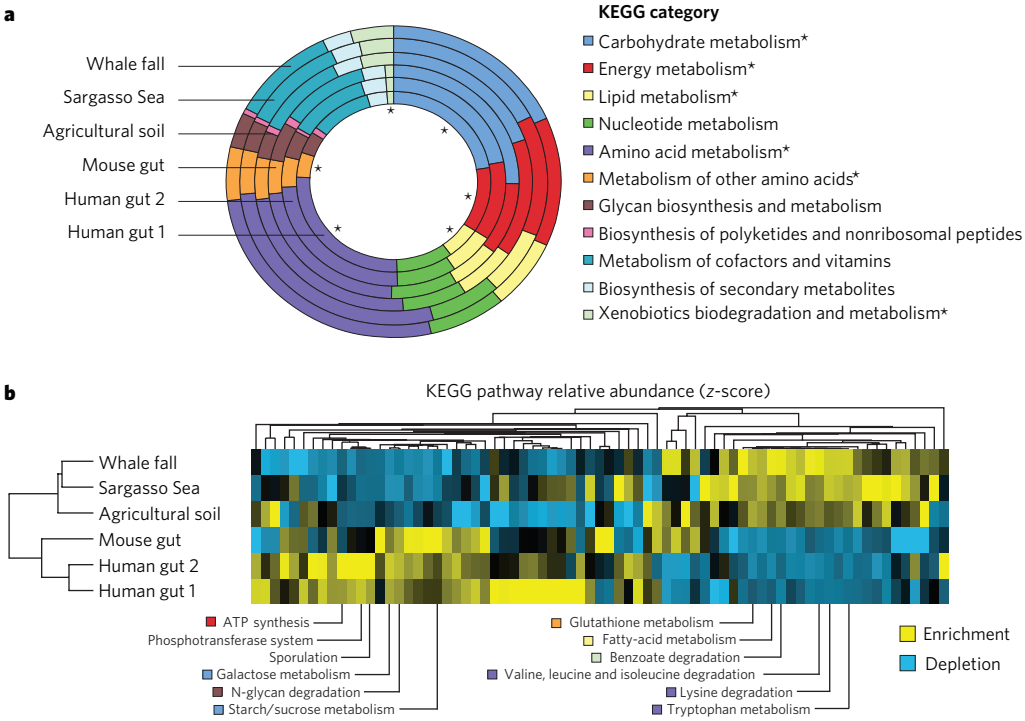


Figure 2 | Functional comparison of the gut microbiome with other sequenced microbiomes. a, Relative abundance of predicted genes, assigned to KEGG categories for metabolism. Several gut-microbiome data sets were analysed: a combined mouse gut data set ($n = 5$ animals)[19] and two human gut data sets[1]. Three 'environmental'-microbiome data sets were also analysed: a combined whale-fall data set ($n = 3$ samples, from three separate whale falls)[20], an agricultural soil data set[20] and a combined Sargasso Sea data set ($n = 7$ samples)[21]. Forward DNA-sequencing reads (from a capillary instrument) were culled from each data set and mapped onto reference microbial and eukaryotic genomes from the KEGG database (version 40; BLASTX best-BLAST-hit E value < 10^{-5})[22]. The best BLAST hit was used to assign each sequencing read to a KEGG orthologous group, which was then assigned to KEGG pathways and categories. The distribution of ~15,000 KEGG-category assignments across each of the six data sets was then used to construct two combined data sets of ~45,000 KEGG-category assignments each. Asterisks indicate categories that are significantly enriched or depleted in the combined gut data set compared with the combined environmental data set (χ^2 test, using the Bonferroni correction for multiple hypotheses, $P < 10^{-4}$). **b**, Hierarchical clustering based on the relative abundance of KEGG pathways. Metabolic pathways found at a relative abundance of more than 0.6% (that is, assignments to a given pathway divided by assignments to all pathways) in at least two microbiomes were selected. These relative-abundance values were transformed into z-scores[20], which are a measure of relative enrichment (yellow) and depletion (blue). The data were clustered according to microbiomes and metabolic pathways by using a euclidean distance metric (Cluster 3.0)[40]. The results were visualized by using Java Treeview[41]. The clustering of environmental data sets was consistent irrespective of the distance metrics used, including Pearson's correlation (centred or uncentred), Spearman's rank correlation, Kendall's tau and city-block distance. The 12 most discriminating KEGG pathways are shown (based on the ratio of the mean gut relative abundance to the mean environmental relative abundance). The KEGG category for each metabolic pathway is indicated by coloured squares. Pathway names without corresponding coloured squares include sporulation (which is involved in cell growth and death) and the phosphotransferase system (which is involved in membrane transport). The gut microbiome is enriched for proteins involved in sporulation (reflecting the high relative abundance of Firmicutes) and for pathways involved in importing and degrading polysaccharides and simple sugars.

As is the case for many ecological studies, we must choose between deep sampling of a small number of sites (individual people and body habitats) and broad sampling. Broad sampling would enable the general principles that control community structure and function to be uncovered. However, deep sampling of body habitats from a few individuals is needed to estimate the distribution of species and genes: these estimates, in turn, will allow modelling of the trade-offs between deeper sampling of fewer individuals and shallower sampling of more individuals. Unlike the situation with the International HapMap Project[30], which sought to describe common patterns of genetic variation in humans, there is no baseline expectation for the amount of diversity in different microbial communities, and the development of careful sampling models will be essential for optimizing the use of resources. Also, given the rapid development of new and more massively parallel sequencing technologies, systematic testing will be required to identify ways to maximize sequencing coverage affordably, while maintaining the ability to analyse and assemble genome fragments.

Ultimately, the goal is to associate differences in communities with differences in metabolic function and/or disease. Thus, another key challenge for the HMP is to define the concept of 'distance' between communities and to associate these distances with host biology and various metadata. UniFrac[11,31,32] and other phylogenetic techniques address this problem for 16S rRNA gene data sets and could be extended to the assessment of metagenomic data. With the distances defined, statistical techniques will need to be developed and refined so that multivariate data sets can be integrated into a unified framework, enabling the components of the microbiome that could affect human health and disease to be identified.

The HMP will also require researchers to move beyond comparative genomics to an integrated 'systems metagenomics' approach that accounts for microbial community structure (the microbiota), gene content (the microbiome), gene expression (the 'meta-transcriptome' and 'meta-proteome') and metabolism (the 'meta-metabolome'). Some progress has been made towards generating 'functional gene arrays', to determine the relative abundance of specific genes or transcripts in microbiomes[33–35]. More work is needed to improve the sensitivity of gene arrays and to apply this approach to complex communities such as the human microbiome. The construction and sequencing of complementary DNA libraries form an alternative approach, and these have already been used to examine microbial and eukaryotic mRNA from environmental samples[36,37]. However, high-throughput methods for eliminating highly abundant transcripts (for example, those from rRNA genes) are needed.

Box 2 | A proposal for staging the Human Microbiome Project

In this conceptualization, the HMP is portrayed as a three-tiered effort, with the first tier composed of three components (or pillars).

First tier: initial data acquisition and analysis

Pillar one: construct deep draft assemblies of reference genomes

- Select cultured representatives of microbial divisions in a given habitat by examining 'comprehensive' 16S-rRNA-gene-based surveys
- Create a publicly accessible database of human-associated 16S rRNA gene phylotypes (which could be referred to as the 'virtual microbial body') to facilitate selection by allowing comparisons within and between body habitats, within and between individuals, and between separate studies; and develop faster and better alignment algorithms for building phylogenetic trees
- Obtain phylotypes of interest from existing culture collections (both public and 'private'), with consent to deposit sequence data in the public domain
- Improve technology for culturing organisms that cannot be cultured at present
- Select a subset of 'species' for pan-genomic analysis (that is, the characterization of multiple isolates of a species-level phylotype), and develop better methods for detecting horizontal gene transfer
- Ensure data flow to, and data capture by, the Protein Structure Initiative (http://www.structuralgenomics.org)
- Deposit sequenced isolates, together with information about habitat of origin, conditions for growth and phenotypes, in a public culture repository that can maintain and distribute microorganisms

Pillar two: obtain reference microbiome data sets

- Focus on monozygotic and dizygotic twin pairs and their mothers
- Determine the advantages and disadvantages of different DNA-sequencing platforms
- Characterize, at a preliminary level, within-sample (α) diversity and between-sample (β) diversity
- Ensure the availability of user-friendly public databases in which biomedical and environmental metagenomic data sets are deposited, together with sample metadata
- Develop and optimize tools (distance metrics) for comparing 16S rRNA gene and community metagenomic data sets, and feed back to the pipeline in which cultured or retrieved representatives of different habitat-associated communities are selected and characterized
- Establish specimen and data archives with distribution capabilities
- Generate large-insert microbiome libraries for present and future functional metagenomic screens
- Coordinate with environmental metagenomics initiatives so that efforts to develop resources and tools are reinforced and shared

Pillar three: obtain shallower 16S rRNA gene and community metagenomic data sets from moderate number of samples

- Extend sampling of families (for example, to fathers, siblings and children of twins), expand the age range of individuals sampled, and explore demographic, socio-economic and cultural variables
- Establish a global sample-collection network, including countries in which social structures, technologies and lifestyles are undergoing rapid transformation
- Develop and optimize computational tools and metrics for comparing these diverse multivariate data sets
- Develop and optimize tools for analysing the transcriptome, proteome and metabolome, by using the same biological specimens used for sequencing community DNA, and develop and optimize tools for higher-throughput analyses
- Design and test experimental models for identifying the principles that control the assembly and robustness of microbial communities

Second tier: choice of individuals that represent different clusters, for additional deep sequencing

- Estimate sampling depth and number of individuals needed to characterize the 'full' human microbiome; the granularity of the characterization needs to match the data
- Search for relatives of human-associated microbial species and gene lineages in other mammalian microbial communities and in the environment, and sequence the genomes of these microorganisms (defining niches; feed back to the first tier)

Third tier: global human microbiome diversity project

- Sequence at a shallow level the microbiomes from a large (to be defined) sample of geographically, demographically and culturally diverse individuals
- Choose individuals with different clinical 'parameters', and carry out association studies and biomarker panning
- Sequence at a large scale reservoirs of microorganisms and genes (for example, soils and water sources), and associate this information with the fluxes of energy, materials, genes and microbial lineages into the human microbiome (with the help of microbial observatories and human observatories)
- Apply the knowledge gained (for example, towards developing diagnostic tests, therapies and strategies for improving the global food chain), and educate people (including the public, governments, and present and future researchers in the field)

Proteomic tools, including Elucidator (http://www.rosettabio.com/products/elucidator) and SEQUEST (http://fields.scripps.edu/sequest), are also available for analysing complex samples. And comprehensive microbial protein-sequence databases (for example, Protein Clusters; http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters) are continually updated. In addition, custom databases can be created from metagenomic data sets and used to interpret mass-spectrometry data sets[38]. Given the limited knowledge of the biological transformations that human microbial communities support, meta-metabolomics is likely to be challenging. Tools and databases for metabolite identification still need to be developed, despite the existence of highly accurate instrumentation. (For example, Fourier-transform ion-cyclotron-resonance mass spectrometers have a mass accuracy of < 1–10 parts per million.) This situation should be helped by ambitious efforts that are underway to catalogue thousands of human-associated metabolites and to generate a searchable database[39]. Together, these complementary measurements will allow a far richer characterization of human microbial communities. They will also enable the variation that is typical of a healthy state to be defined, making it possible to search for deviations that are associated with disease.

Depositing and distributing data

Vast amounts of information will be generated by the HMP, as well as by metagenomic surveys of the environment, so new procedures and increased capabilities are required for depositing, storing and mining different data types. Important goals include the following: a minimum set of standards for annotation; a flexible, simple and open format for depositing metadata (taking a lesson from clinical studies because the relevant parameters are largely unknown); efficient analysis tools for the general user that are broadly applicable (including tools for meta-analyses of varied data types); and an adequate cyberinfrastructure to support the computing needs of the research community.

Using model systems

Although the HMP is human-focused, model organisms and other experimental systems are needed for aspects of the project that cannot be tested in humans: these will define how communities operate and interact with their hosts, characterize the determinants of community robustness and identify biomarkers of community composition and/or performance. Germ-free animals, both wild-type and genetically engineered, that have been colonized at various stages of their lives with simplified microbial communities composed of a few sequenced members, or with more complex consortia, should be useful because they provide the opportunity to constrain several variables, including host genotype, microbial diversity and environmental factors such as diet. *In vitro* models, including microfluidic-based techniques for single-cell sorting and measurements, should help to define the biological properties of microorganisms and the consequences of interactions between microorganisms.

A model for staging the HMP

On the basis of all of these considerations, one potential way of staging the HMP is outlined in Box 2. The search for data will be global in many senses. It embraces the planet and its (human) inhabitants. It requires individuals from the clinical, biological and physical-engineering sciences to participate, including those with expertise in disciplines ranging from mathematics to statistics, computer science, computational biology, microbiology, ecology, evolutionary biology, comparative genomics and genetics, environmental and chemical engineering, chemistry and biochemistry, human systems physiology, anthropology, sociology, ethics and law. It requires coordination between scientists, governments and funding agencies. And it is one element of a worldwide effort to document, understand and respond to the consequences of human activities — not only as they relate to human health but also as they relate to the sustainability of the biosphere. It is hoped that, just as microbial observatories have been set up to monitor changes in terrestrial and ocean ecosystems worldwide, an early outcome of the HMP will be the establishment of 'human observatories' to monitor the microbial ecology of humans in different settings.

Concluding remarks

Many outcomes of the HMP can be predicted: for example, new diagnostic biomarkers of health, a twenty-first century pharmacopoeia that includes members of the human microbiota and the chemical messengers they produce, and industrial applications based on enzymes that are produced by the human microbiota and can process particular substrates. One important outcome is anticipated to be a deeper understanding of the nutritional requirements of humans. This, in turn, could result in new recommendations for food production, distribution and consumption that are formulated based on knowledge of the microbiome. ■

Peter J. Turnbaugh, Ruth E. Ley and Jeffrey I. Gordon are at the Center for Genome Sciences, Washington University School of Medicine, St Louis, Missouri 63108, USA. Micah Hamady is at the Department of Computer Science, University of Colorado at Boulder, Boulder, Colorado 80309, USA. Claire M. Fraser-Liggett is at the Institute of Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. Rob Knight is at the Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, Colorado 80309, USA.

1. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
2. The Committee on Metagenomics. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (The National Academies Press, Washington DC, 2007).
3. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton, 1967).
4. Ambramsky, Z. & Rosenzweig, M. L. The productivity diversity relationship: Tilman's pattern reflected in rodent communities. *Nature* **309,** 150–151 (1984).
5. Ley, R. E. *et al.* Obesity alters gut microbial ecology. *Proc. Natl Acad. Sci. USA* **102,** 11070–11075 (2005).
6. Rawls, J. F., Mahowald, M. A., Ley, R. E. & Gordon, J. I. Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127,** 423–433 (2006).
7. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32,** 1363–1371 (2004).
8. Cole, J. R. *et al.* The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33,** D294–D296 (2005).
9. Schloss, P. D. & Handelsman, J. DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71,** 1501–1506 (2005).
10. DeSantis, T. Z. *et al.* NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* **34,** W394–W399 (2006).
11. Lozupone, C., Hamady, M. & Knight, R. UniFrac — an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* **7,** 371 (2006).
12. Gao, Z. *et al.* Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl Acad. Sci. USA* **104,** 2927–2932 (2007).
13. Pei, Z. *et al.* Bacterial biota in the human distal esophagus. *Proc. Natl Acad. Sci. USA* **101,** 4250–4255 (2004).
14. Bik, E. M. *et al.* Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl Acad. Sci. USA* **103,** 732–737 (2006).
15. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
16. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: Human gut microbes associated with obesity. *Nature* **444,** 1022–1023 (2006).
17. Hyman, R. W. *et al.* Microbes on the human vaginal epithelium. *Proc. Natl Acad. Sci. USA* **102,** 7952–7957 (2005).
18. Hubbell, S. P. Neutral theory and the evolution of ecological equivalence. *Ecology* **87,** 1387–1398 (2006).
19. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
20. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308,** 554–557 (2005).
21. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
22. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277–D280 (2004).
23. Gordon, J. I. *et al.* Extending our view of self: the Human Gut Microbiome Initiative (HGMI). *National Human Genome Research Institute* <http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf (2005).
24. Xu, J. *et al.* Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biol.* **5,** e156 (2007).
25. Podar, M. *et al.* Targeted access to the genomes of low abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73,** 3205–3214 (2007).
26. Zengler, K. *et al.* Cultivating the uncultured. *Proc. Natl Acad. Sci. USA* **99,** 15681–15686 (2002).
27. Teeling, H. *et al.* TETRA: a web-service and stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5,** 163 (2004).

28. McHardy, A. C. *et al.* Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* **4,** 63–72 (2007).
29. von Mering, C. *et al.* Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315,** 1126–1130 (2007).
30. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
31. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).
32. Lozupone, C. A., Hamady, M., Kelley, S. T. & Knight, R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.* **73,** 1576–1585 (2007).
33. Wu, L. *et al.* Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.* **67,** 5780–5790 (2001).
34. Gentry, T. J. *et al.* Microarray application in microbial ecology research. *Microb. Ecol.* **52,** 159–175 (2006).
35. Gao, H. *et al.* Microarray-based analysis of microbial community RNAs by whole-community RNA amplification. *Appl. Environ. Microbiol.* **73,** 563–571 (2007).
36. Poretsky, R. S. *et al.* Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* **71,** 4121–4126 (2005).
37. Grant, S. *et al.* Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. *Appl. Environ. Microbiol.* **72,** 135–143 (2006).
38. Ram, R. J. *et al.* Community proteomics of a natural microbial biofilm. *Science* **308,** 1915–1920 (2005).
39. Wishart, D. S. *et al.* HMDB: the human metabolome database. *Nucleic Acids Res.* **35,** D521–D526 (2007).
40. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20,** 1453–1454 (2004).
41. Saldanha, A. J. Java Treeview — extensible visualization of microarray data. *Bioinformatics* **20,** 3246–3248 (2004).
42. Backhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101,** 15718–15723 (2004).
43. Backhed, F., Manchester, J. K., Semenkovich, C. F. & Gordon, J. I. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. *Proc. Natl Acad. Sci. USA* **104,** 979–984 (2007).
44. Martin, F. J. *et al.* A top-down systems biology view of microbiome–mammalian metabolic interactions in a mouse model. *Mol. Syst. Biol.* **3,** doi:10.1038/msb4100153 (2007).
45. Sidhu, H., Allison, M. J., Chow, J. M., Clark, A. & Peck, A. B. Rapid reversal of hyperoxaluria in a rat model after probiotic administration of *Oxalobacter formigenes*. *J. Urol.* **166,** 1487–1491 (2001).
46. Chu, F. F. *et al.* Bacteria-induced intestinal cancer in mice with disrupted *Gpx1* and *Gpx2* genes. *Cancer Res.* **64,** 962–968 (2004).
47. Pull, S. L., Doherty, J. M., Mills, J. C., Gordon, J. I. & Stappenbeck, T. S. Activated macrophages are an adaptive element of the colonic epithelial progenitor niche necessary for regenerative responses to injury. *Proc. Natl Acad. Sci. USA* **102,** 99–104 (2005).
48. Hooper, L. V., Stappenbeck, T. S., Hong, C. V. & Gordon, J. I. Angiogenins: a new class of microbicidal proteins involved in innate immunity. *Nature Immunol.* **4,** 269–273 (2003).
49. Mazmanian, S. K., Liu, C. H., Tzianabos, A. O. & Kasper, D. L. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122,** 107–118 (2005).
50. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
51. Cash, H. L., Whitham, C. V., Behrendt, C. L. & Hooper, L. V. Symbiotic bacteria direct expression of an intestinal bactericidal lectin. *Science* **313,** 1126–1130 (2006).
52. Braun-Fahrlander, C. *et al.* Environmental exposure to endotoxin and its relation to asthma in school-age children. *N. Engl. J. Med.* **347,** 869–877 (2002).
53. Kozyrskyj, A. L., Ernst, P. & Becker, A. B. Increased risk of childhood asthma from antibiotic use in early life. *Chest* **131,** 1753–1759 (2007).
54. Wostmann, B. S., Bruckner-Kardoss, E. & Pleasants, J. R. Oxygen consumption and thyroid hormones in germfree mice fed glucose–amino acid liquid diet. *J. Nutr.* **112,** 552–559 (1982).

Acknowledgements We apologize that we could not cite many excellent studies because of space constraints.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to J.I.G. (jgordon@wustl.edu).

An ecological and evolutionary perspective on human–microbe mutualism and disease

Les Dethlefsen[1], Margaret McFall-Ngai[2] & David A. Relman[1,3,4]

The microbial communities of humans are characteristic and complex mixtures of microorganisms that have co-evolved with their human hosts. The species that make up these communities vary between hosts as a result of restricted migration of microorganisms between hosts and strong ecological interactions within hosts, as well as host variability in terms of diet, genotype and colonization history. The shared evolutionary fate of humans and their symbiotic bacteria has selected for mutualistic interactions that are essential for human health, and ecological or genetic changes that uncouple this shared fate can result in disease. In this way, looking to ecological and evolutionary principles might provide new strategies for restoring and maintaining human health.

Nowhere in the study of human biology are basic concepts changing more rapidly than with respect to the human microbiota. Microorganisms were first shown to cause disease in humans in the 1800s, and after this finding, the popular and scientific views of the microbial world became dominated by the quest to understand, prevent and cure microbial disease. This led to millions of lives being saved through improved hygiene, vaccinations and antibiotics. However, most interactions between humans and microorganisms do not result in disease. Beneficial host–microbe interactions have been studied for more than a century, but it was not until the advent of molecular biology that the pathogen-dominated view of human-associated microorganisms began to change. Gene-sequence-based approaches have recently allowed complex microbial communities to be characterized more comprehensively and have removed the constraint of being able to identify only microorganisms that can be cultured, greatly increasing knowledge about commensal microorganisms and mutualistic microorganisms of humans[1–12] (that is, organisms in a relationship in which one partner benefits and the other is unharmed, and organisms in a relationship in which both partners benefit, respectively), as well as human pathogens[13–18]. Researchers are now finding that host–microbe interactions are essential to many aspects of normal 'mammalian' physiology, ranging from metabolic activity to immune homeostasis[19–25]. With the availability of new tools to investigate complex microbial communities and the expanded appreciation for the importance of the human indigenous microbiota, this is an opportune time to apply ecological and evolutionary principles to improve the current understanding of both health and disease.

So far, the human microbiota has not been fully described, but it is clear that microorganisms are present in site-specific communities on the skin and mucosal surfaces and in the intestinal lumen. Each community contains microorganisms from certain families and genera that are found in the same habitat in many or most individuals, although at the species and strain levels the microbiota of an individual can be as unique as a fingerprint[3,11,26]. The microbial communities of other terrestrial vertebrates mainly contain lineages that are related to, but distinct from, those in humans[27–31]. These characteristics indicate that humans have co-evolved with their microbial partners. In this review, we examine evolutionary and ecological principles that are relevant to these relationships, and we consider microbial pathogenesis in this context.

Evolution of mutualism

In the 1960s, evolutionary biologists rejected the idea that natural selection would generally favour the good of the species (or any group), because individual types with the greatest reproductive success in a population increase in relative abundance regardless of the consequences for the population as a whole[32]. Since then, the evolution of traits that benefit individuals other than the trait bearer has been extensively researched, both theoretically and empirically. Although the field has been contentious at times, there is now general agreement about the conditions that promote cooperation, including mutualism between species[32–34].

Organisms can have traits that contribute directly to their own fitness and also incidentally benefit members of another species. When such 'by-product benefits' occur in both directions, the result is a no-cost mutualism[34]. For example, plant polysaccharides that are not digestible by humans are the main substrates for microbial growth in the colon, whereas butyrate and other products of microbial fermentation are important energy sources for the host[35,36]. Intestinal symbionts are selected to be effective consumers of available resources through direct effects on their fitness, but this also benefits the host because resource competition provides an additional barrier to colonization by potential pathogens[37–41].

If mutualistic by-product interactions such as the example above are possible, but not ecologically inevitable, then traits that improve the likelihood or stability of a relationship (for example, site-specific attachment molecules) might evolve in one or both partners. A species might also evolve to increase its own fitness by increasing the fitness of a mutualistic partner[34]. For example, microbial symbionts that secrete molecules that inhibit host pathogens (known as pathogen interference)[38–40] or detoxify compounds that harm the host[42] can augment the lifespan and reproductive capacity of the host, thereby giving the symbionts more opportunities to spread. Evolution to increase mutualistic benefits has been called 'partner fidelity feedback', and it is strengthened if the same lineages of partners interact across multiple generations[33,34]. Unlike traits that support mutualism incidentally, traits that evolve specifically to improve

[1]Department of Microbiology and Immunology, Stanford University, Stanford, California 94305, USA. [2]Department of Medical Microbiology and Immunology, Symbiosis Cluster, 4835A Medical Sciences Center, 1300 University Avenue, University of Wisconsin, Madison, Wisconsin 53706, USA. [3]Department of Medicine, Stanford University, Stanford, California 94305, USA. [4]Veterans Affairs Palo Alto Health Care System 154T, 3801 Miranda Avenue, Palo Alto, California 94304, USA.

mutualism, such as the production of compounds dedicated to pathogen interference, can impose a direct fitness cost, although a net benefit would be expected in the context of the evolved mutualism[34].

Mutualism-promoting traits with a direct cost for the bearer, however, create the potential for 'cheating'. When organisms interact to create a shared benefit, cheaters are organisms that obtain the benefit without helping to create it. For example, a cheating microbial phenotype could result from a mutation that redirects resources towards faster growth of the microorganism itself instead of detoxification or pathogen interference. The cheater therefore increases its relative fitness in a host by avoiding costly contributions towards host fitness while benefiting from the improved host fitness that results from the mutualistic contributions of its competitors[33,34]. Various evolutionary outcomes are possible, including the absence of costly contributions to mutualism, contributions to mutualism that are below the level that would maximize the mutualistic benefit, and the coexistence of mutualists and cheaters in the community[43]. A possible example of this dynamic balance is that certain benefits attributed to probiotic bacterial species are characteristic of only a subset of the strains that make up the species[38,40]. For any mutualism that is not cost free, the partners can evolve mechanisms to protect their relationship from being exploited by cheaters[32,34,44], and mutualism can be stronger and more stable where ecological features limit the potential for exploitation (discussed later).

The immune system is the most conspicuous set of anti-exploitation adaptations involved in human–microbial symbiosis. The gastrointestinal mucosa is intimately associated with the most abundant and diverse microbial communities in the human body, but the gut-associated immune system neither controls the composition of the gut microbiota nor remains ignorant of it. Instead, specialized tissues and cells actively sample the intestinal contents and initiate local immune responses that help to confine the microbiota to the gut, avoiding a damaging systemic inflammatory response to the microorganisms present in the healthy gut[45]. However, if host tissue is damaged[46] or if microorganisms spread to normally sterile sites[45], then there is a vigorous systemic response to clear the infection. Therefore, microorganisms are free to compete for resources in the gut, generating a robust and disease-resistant community[37,41], but are prevented (usually) from exploiting the host to obtain additional resources. Recent work has also shown that the normal development and activity of the 'host' immune system is itself a result of mutualistic interactions[20–22,24,25] (see page 819).

Humans and their collective microbiota are segmented into many local communities, each comprising an individual human with his or her symbionts. This ecological pattern, characterized by strong interactions within distinct local communities and limited interactions or migration between them, is described as a metacommunity. Another level of metacommunity organization exists because individual humans belong to social groups that tend to share a similar microbiota[47,48]. At both levels, the metacommunity structure allows selection to occur between the local units (that is, between individuals or between social groups), which promotes mutualism and restrains cheating within the human–microbe symbiosis[32,49]. Such selection occurs when a local symbiotic community succeeds or fails together, with more successful communities increasing in abundance or prevalence relative to less successful communities[32]. For example, a human individual or social group that carries a microbiota with strong defences against an abundant pathogen is likely to leave more progeny than those lacking such defences. If the progeny tend to carry the parental microbiota, then mutualistic microorganisms that make costly contributions to pathogen defence are favoured by selection between distinct local symbiotic communities. This community-level selection opposes the tendency for cheating non-defenders to increase in relative abundance within each local symbiotic community[32]. The greater similarity of the microbiota within a human family than between human families[12] (and within, rather than between, chimpanzee social groups[30]) shows that there is, indeed, a shared evolutionary fate. The individualized microbiota of each person has a stake in his or her fitness.

Human–microbe mutualism often involves more than two partners, although the same principles apply. For example, the colonic degradation of polysaccharides that provides butyrate for the host is a cooperative microbial process[35,36]. Extracellular enzymes from multiple species are required for complete hydrolysis of the polymers. In addition, some of the resultant sugars are consumed by strains that do not produce extracellular enzymes but provide growth factors to strains that do[35]. Some fermenters such as *Bifidobacterium* spp. release lactate as waste. Their fermentation efficiency is increased by lactate fermenters, such as *Eubacterium hallii*, that release butyrate as waste, and this butyrate is then used by the host[36]. Sugar-fermenting lactobacilli that produce neither hydrolytic enzymes nor growth factors could be considered cheaters from the perspective of polysaccharide degradation, but they could be considered mutualists of the entire symbiotic community if they interfere with the colonization of pathogens[40]. The butyrate-producing consortium as a whole is a mutualist of the host and would be favoured by community-level selection over consortia producing less-desirable fermentation products[36]. However, selection for mutualistic functional traits such as butyrate production cannot entirely determine the composition of the microbiota, because communities of different composition can have similar functional characteristics in a given context. Not only selection on community-level traits but also competition within the community and chance colonization events affect the structure of the microbiota[50].

Human microbial communities and health

The distribution of microorganisms in and on the human body reflects adaptations to life on land, which were made about 400 million years ago. Terrestrial vertebrates developed skin, lungs, internal fertilization, and protective membranes around the embryo. The skin became relatively impermeable, and mucous membranes were confined to protected sites. Because microorganisms generally thrive only in moist environments, these adaptations to a mostly dry environment have shaped the abundance, location and phenotypes of human-associated microorganisms and have limited the exchange of microorganisms between individuals.

Table 1 | Model systems for animal–microbe symbioses

Type of symbiosis	Specific system (Host/symbiont species)	Host phylogenetic affiliation	Host tissue colonized	Reference
Highly complex consortia (10^2–10^3)*	*Mus musculus* (mouse)	Vertebrate chordate	Intestine	19
	Danio rerio (zebrafish)	Vertebrate chordate	Intestine	86
	Microcerotermes spp. and *Reticulitermes* spp. (termites)	Insect arthropod	Hindgut	87
Relatively simple consortia (~2–25)*	*Hirudo medicinalis* (leech)	Oligochaete annelid	Intestine	88
	Lymantria dispar (gypsy moth)	Insect arthropod	Larval midgut	89
	Drosophila melanogaster (fruitfly)	Insect arthropod	Intestine	90
	Hydra oligactis and *Hydra vulgaris*	Hydrozoan cnidarian	Not determined	91
Monospecific (1)*	*Euprymna scolopes* (sepiolid squid)/*Vibrio fischeri*	Cephalopod mollusc	Light organ	92
	Eisenia fetida (earthworm)/*Acidovorax* spp.	Oligochaete annelid	Excretory tissues	93
	Steinernema spp./*Xenorhabdus* spp. and *Heterorhabditis* spp./*Photorhabdus* spp.	Entomopathogenic nematodes	Gut-associated vesicle or region	94

*Number of bacterial-symbiont phylotypes found reproducibly.

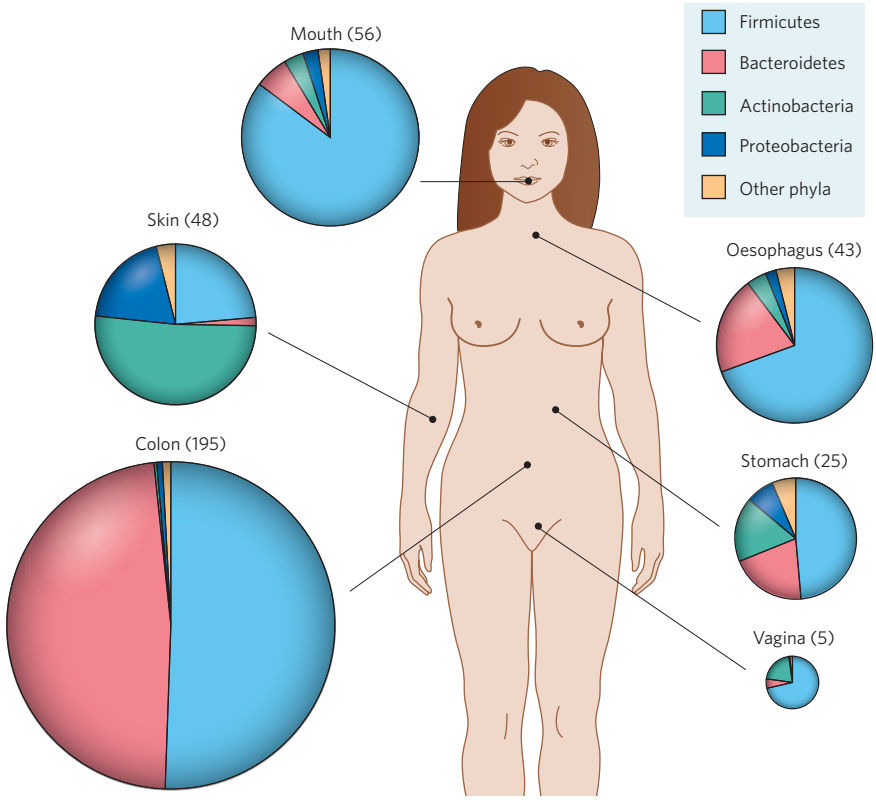


Figure 1 | Site-specific distributions of bacterial phyla in healthy humans. The area of the chart for each site represents the average number of distinct phylotypes (approximate species-level taxa, based on 16S rRNA gene-sequence analysis) per individual. (The mean number of phylotypes per individual is shown in parentheses; 3–11 individuals were studied per habitat.) The coloured wedges represent the proportion of phylotypes belonging to different phyla. More than 50 bacteria phyla exist, but human microbial communities are overwhelmingly dominated by the 4 that are shown. The relative abundance of these phyla at most sites tends to be consistent across individuals: for example, in almost all humans studied so far, Bacteroidetes and Firmicutes predominate in the colon. By contrast, the composition of the vaginal microbiota is more variable; most women have a preponderance of Firmicutes with few other representatives, whereas a minority of women have a preponderance of Actinobacteria with few other representatives. An estimated 20–80% of human-associated phylotypes (depending on habitat) are thought to have eluded cultivation so far. Data taken from refs 1–7.

The current understanding of the human microbiota relies heavily on cultivation-based approaches and therefore is biased and incomplete. Although imperfect, molecular approaches that identify microorganisms from small-subunit (16S) ribosomal RNA gene sequences offer advantages over cultivation. The 16S rRNA gene is typically chosen because it is present universally and can provide a taxonomic identification ranging from the domain and phylum level to approximately the species level. However, these methods have been used to study human microbial ecology for only a decade, and the available data are limited. There are few deep surveys of microbial-community membership in any human habitat and even fewer assessments of functional potential or activity. In general, 16S rRNA gene-sequence data have been collected from one site in a few humans at one time, representing a narrow range of health and disease states[2,5–7], although there are studies that include several temporal or spatial samples per individual[1,3,4,51]. Sequence-dependent approaches that are less labour-intensive but yield lower-resolution data have been applied to a larger number of individuals, at various time points and under various conditions[8,9,11,12,52]. Even so, the microbial communities associated with only a small proportion of the diversity of human genotypes, lifestyles, diets and diseases have been investigated. One high-throughput method for obtaining information about bacterial communities is to use phylogenetic microarrays, which yield high-resolution data, but this method also depends on adequate 16S rRNA gene-sequence databases[10]. Like these microarray studies, metagenomic and proteomic analyses are just beginning to be published[53,54]. Technical and ethical constraints restrict sampling from humans; therefore, model systems will continue to be important, and examples of these are listed in Table 1.

Despite the limited data available, analyses of the human microbiota have revealed intriguing features. Most of the phylogenetic diversity is found in shallow, wide radiations in a small subset of the known deep lineages[26]. Specifically, there are more than 50 bacterial phyla on Earth, but human-associated communities are dominated by only 4 phyla (Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria), with 9 other phyla (Chlamydiae, Cyanobacteria, Deferribacteres, Deinococcus–Thermus, Fusobacteria, Spirochaetes, Verrucomicrobia, and the candidate phyla TM7 and SR1) found in some sites and individuals (Fig. 1). In contrast to the paucity of phyla represented, the human microbiota contains an abundance of species and strains. Uniform probabilities of speciation and extinction over time would result in an exponential increase in the number of lineages throughout evolution. However, in humans, there is a marked excess of phylotype diversity at the species and strain level compared with the trends in more inclusive taxa (Fig. 2); there are similar patterns in other vertebrate hosts[26]. This finding might reflect a long history of stability in the types of microbial niche associated with terrestrial animals, together with factors (such as host heterogeneity and metacommunity structure) that promote diversification among organisms in similar niches. In contrast to the remarkable diversity of bacterial species, a striking but unexplained finding is that the only Archaea found frequently in humans are several species of methanogens. *Methanobrevibacter smithii* is abundant in the colon of some humans[3,53]. Also, *Methanobrevibacter oralis* and close relatives can be found within the subgingival crevice in the human mouth but only in the setting of moderate to severe disease[55]. Overall, the human microbiota is similar to that of other mammals at the phylum level, but most bacterial families and genera seem to be distinct (Fig. 3).

Multiple samples of the microbiota that are taken separated in time or space from a single body habitat within one individual are generally more similar to each other than they are to samples from the same habitat in another individual[3,9,11], although temporal variation in the skin microbiota of an individual is as great as the variation between individuals[4]. In addition, the bacterial communities at a given site are more similar between human family members than between unrelated individuals[12], but more studies are necessary to distinguish the effects of genetic relatedness and a shared early environment[50]. Antibiotics can markedly affect the composition of the microbiota in the short term, with most (but not all) families and genera of gut microorganisms returning to typical levels within weeks of exposure[51,56]. However, pathogens can exploit the reduced competitiveness of a community disturbed by antibiotics, thereby establishing themselves in the host[39,57]. The degree to which the unique bacterial communities of an individual are re-established after antibiotic treatment is unclear, but particular antibiotic-resistant strains that colonize or evolve during treatment can persist for years[58,59].

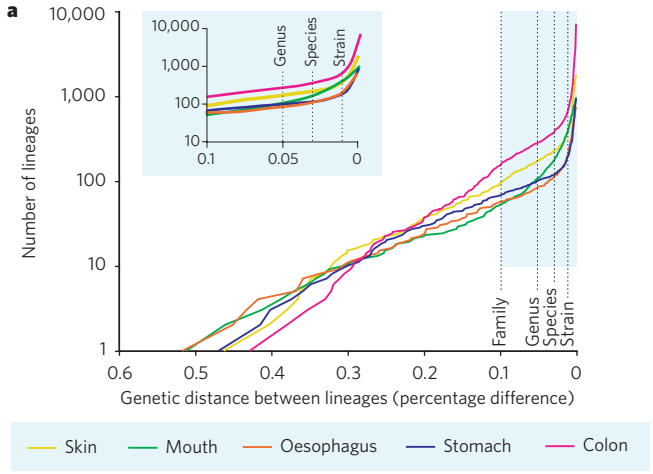


Figure 2 | Patterns of human-associated microbial diversity. a, Lineage-by-distance analysis of 16S rRNA gene-sequence data from human microbial communities in specific habitats. The *x* axis shows the percentage difference threshold (Olsen correction), over 1,241 unambiguously aligned positions of near full-length 16S rRNA gene sequences, for delineating separate lineages. The *y* axis shows the number of distinct lineages that exist at the distance threshold. If speciation and extinction occur with constant probabilities as 16S rRNA gene sequences diverge, this would result in an exponentially increasing number of lineages with diminishing evolutionary distances between them (a straight line on a semi logarithmic plot). Such a pattern seems to hold from the phylum level (largest distances between lineages) to approximately the species level. However, relative to this trend, all sites have an excess of recently diverged lineages. The excess lineages accumulate in the range of 16S rRNA gene divergence that is typically associated with species and strains. The inset depicts a portion of the same data at a larger scale. Samples were taken from 3–11 individuals, depending on the site. Data taken from refs 1–5. **b**, When displayed as a dendrogram, 16S rRNA gene-based patterns of microbial diversity in soil and aquatic environments generally resemble the tree shape on the left, with new branches arising at all distances from the root. Patterns of diversity in vertebrate-associated communities resemble the tree shape on the right, with few branches arising close to the root and many branches arising close to the branch tips.

A human infant acquires its microbiota from the environment. In humans, symbionts are not vertically transmitted (that is, transmitted through the germ line), as they are in some invertebrate animals. Colonization, succession and diversification occur within characteristic windows of time in the various microbial habitats in the body, ranging over the first weeks, months or years of life[10,60,61]. The composition of faecal communities in early infancy, for example, is dynamic and reflects opportunistic environmental exposures, especially to the mother. The introduction of solid foods then begins the transition to an individualized, adult-like microbiota[10]. The assembly of bacterial communities on tooth surfaces also follows a consistent pattern as teeth emerge[62], as well as after the removal of pre-existing biomass (that is, plaque)[52,60].

The horizontal transfer of microorganisms to every human generation favours strains that are locally abundant at that time (for example, those present in parents and kin), but colonization remains somewhat stochastic[50]. The mixing of lineages from different sources that occurs during community assembly is analogous to the reassortment of parental alleles during sexual reproduction[49], and it promotes the adaptation of community composition to local conditions and the rapid spread of beneficial strains. However, strains that become locally abundant by cheating can also spread.

Competition for niches within the human microbiota is ubiquitous and occurs together with the selective forces that promote mutualism in the community as a whole. Microorganisms can even compete and cooperate simultaneously. For example, *Bacteroides thetaiotaomicron* and *M. smithii* facilitate each other's growth by complementary energy metabolism, while competing for nitrogen[63]. Cooperatively crosslinked biofilms containing multiple species promote the colonization of tooth surfaces, even while the constituent species compete with each other for individual binding sites[64]. Symbionts that are highly prevalent and abundant probably have effective mechanisms for competing for resources: for example, *B. thetaiotaomicron*[65] and *Bifidobacterium longum*[66] have a wide variety of inducible genes encoding factors involved in the binding, uptake and degradation of plant- and host-derived polysaccharides. Competition within the human microbiota involves not only resources but also interference; that is, the direct inhibition of one strain by another in a resource-independent manner. In some cases, the metabolic by-products of one species (such as lactate or short-chain fatty acids) inhibit other microorganisms. In other cases, dedicated compounds are generated solely because of their inhibitory effect: for example, reactive oxygen species and the peptide antibiotics known as bacteriocins[39,40]. The immediate fitness costs and context-dependent benefits of dedicated interference compounds result in selection for diversity: for example, the capacity to produce and resist bacteriocins evolves rapidly among closely related strains[67,68]. Both resource competition and interference competition contribute to the resistance of the intact microbiota to colonization by pathogens[37–40].

Microbial evolution and human disease

Microbial symbionts occupy a complex adaptive landscape. Many traits affect fitness, and many different trait combinations can generate a local optimum fitness (that is, a fitness peak). Natural selection generally acts on subtle phenotypic differences to move microorganisms 'uphill' towards a fitness peak, but larger changes can move an organism onto the slope of a different fitness peak (Fig. 4). The fitness of a symbiont depends on environmental features that can change, such as the coexisting microbiota, the diet of the host, and which species and even particular individual is the host. Thus, the adaptive landscape is dynamic.

Changes in the genotype or environment of a non-pathogenic symbiont can result in the invasion of host tissue. The usual outcome is then an immune response that eliminates the infection. This high rate of microbial mortality imposes a strong selective pressure: the rare changes that enable a symbiont to survive such a challenge would involve avoiding immune recognition or circumventing immune control, at least until some progeny have been transmitted to a new host. Alternatively, changes that increase the opportunities for a non-pathogen to be transmitted to a new host reduce the dependence of the microorganism on the fate of the current host. In this case, the selective pressures on the fitness of the symbiont are less constrained by the need to preserve host fitness as well. In either case, the microorganism can begin adapting towards a fitness peak as a pathogen.

All human microbial symbionts must be able to establish themselves in new hosts. The adaptations of mutualistic or commensal microorganisms towards this end can facilitate a pathogenic lifestyle as well. For example, the biochemical mechanisms for sensing host environments, interacting with host surfaces and even communicating with the host are often the same in human pathogens, commensal microorganisms and mutualistic microorganisms[15,69–71]. Not surprisingly, many common human pathogens are closely related to non-pathogenic symbionts: examples are found in

the genera *Staphylococcus*, *Streptococcus*, *Neisseria* and *Enterococcus*, and in the family Enterobacteriaceae[13–18]. It is not coincidental that these taxa tolerate the aerobic environment between hosts, whereas the more abundant, but less aerotolerant, taxa of the colon have fewer known pathogens as close relatives. The greater ability of aerotolerant taxa to be transmitted to a new host weakens the selection for mutualism in the current host[33,34]. In general, the pathogenic phenotypes in taxa that contain abundant non-pathogenic symbionts have multiple evolutionary origins[13–18], emphasizing that pathogenicity is not necessarily a considerable evolutionary barrier for microorganisms. By contrast, other pathogens have originated only once[72,73], but the continued emergence of new diseases is a reminder that there might be many unoccupied pathogen fitness peaks at present.

The evolution of pathogen virulence has received considerable attention, largely centred on the paradox that pathogens both harm and depend on their hosts[74]. The view that highly virulent pathogens originated recently, with selection inevitably reducing virulence over time, has been supplanted by the realization that there is an optimal level of virulence (for the pathogen) that depends on the biology of its host interactions[74,75]. For example, if pathogen transmission is inherently damaging to the host (as occurs with *Salmonella enterica* serovar Typhimurium[76]), then selective pressure on the pathogen balances the benefit of higher transmission against the loss of host viability as a result of higher virulence. By contrast, pathogens with environmental reservoirs (for example, *Vibrio cholerae*), transmission vectors (for example, *Plasmodium falciparum*) or environmentally resistant propagules (such as spores; for example, *Clostridium tetani*) might be able to afford a higher level of virulence than those that depend on direct transmission[77]. For pathogens that depend on normal host activity for transmission, such as sexually transmitted pathogens (for example, *Chlamydia trachomatis* and *Treponema pallidum*), low virulence and/or long latency can promote the spread of pathogen. In host populations with a reduced potential for pathogens to encounter new hosts, the optimal virulence is reduced to allow the host to survive long enough to ensure pathogen transmission[78].

The observed level of virulence for a pathogen, however, does not necessarily correspond to its evolutionary optimum. Many pathogens are zoonotic (that is, transmitted from animals to humans)[79] and can be adapted to a low-virulence niche in their primary host; an example is enterohaemorrhagic *Escherichia coli* in cattle[80]. Unless transmission by humans contributes to the evolutionary success of the pathogen, excessive (or suboptimal) virulence in humans exerts no selective pressure on the microorganism. Competition between different strains of a pathogen (as a result of co-infection, as occurs with *Plasmodium* spp., or in-host evolution, as occurs with human immunodeficiency virus) can affect virulence, because an optimally virulent pathogen (as measured by transmission success) might not be the best competitor during mixed infections in a single host[81]. A rapidly replicating, excessively virulent strain might kill the host or provoke a successful immune response before transmission of a co-infecting, less virulent strain, even if the latter strain is optimally virulent when infecting a host alone[81]. Competition between pathogens can also decrease virulence. The production of extracellular iron-scavenging molecules (known as siderophores) contributes to the virulence of many bacterial pathogens, but cheating lineages that consume siderophores without producing them reduce virulence, thereby benefiting the host[82]. The diverse biology of host–pathogen and pathogen–pathogen interactions precludes simple predictions about the effect of interpathogen competition on virulence[81,83].

The importance of opportunity for the origin of pathogens is emphasized by a recent analysis of the 25 infectious diseases that cause the most human death and disability[84]. The preferred host of a pathogen is thought to change most easily to a species closely related to the current host[85]. Indeed, although primates constitute only a small proportion of all animal species on Earth, they are the origin of a large proportion of these serious human diseases. However, an even larger proportion of these diseases originated from domestic animals, reflecting greater opportunities for the symbionts of domesticated species to be transmitted to humans[84]. With the advent of agriculture, changes in human populations simultaneously

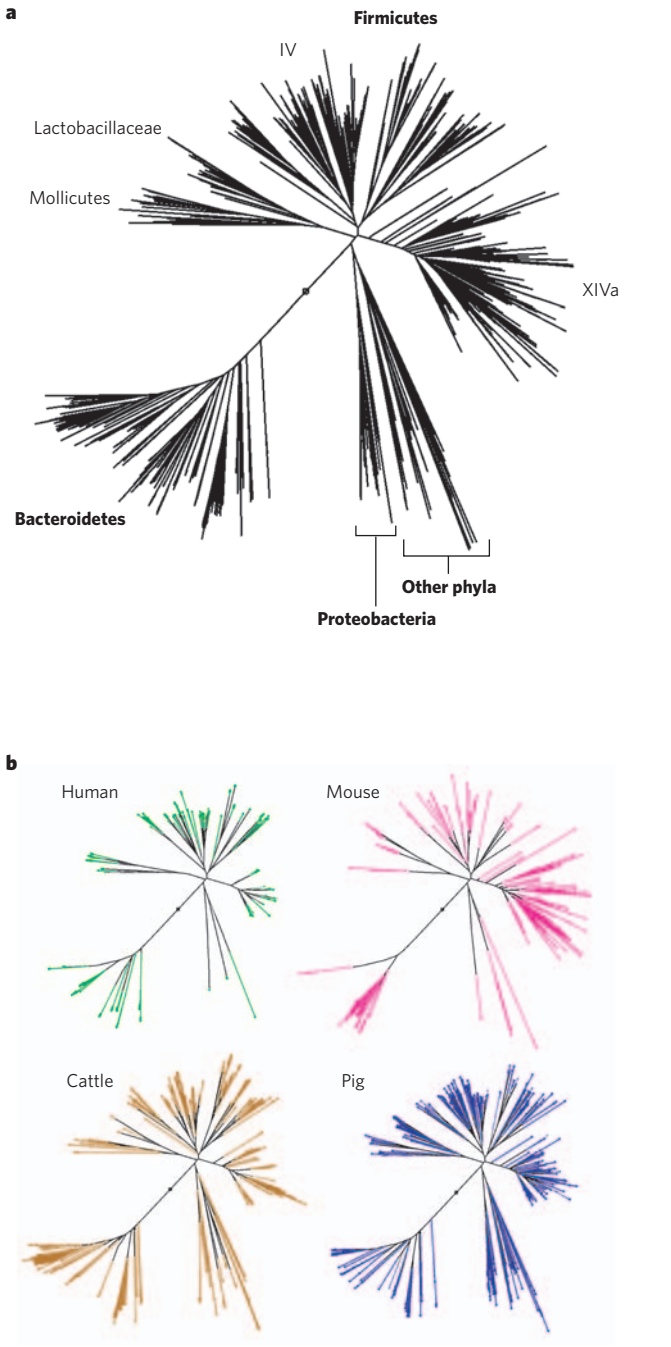


Figure 3 | Relationships between bacterial 16S rRNA gene sequences from the intestinal microbiota of animals. A set of aligned, high-quality, full-length sequences was obtained from Greengenes[95]. Sequences derived from one human stool sample and caecal samples from one mouse family were chosen to obtain approximately the same number of sequences as obtained from multiple studies of the bovine rumen and pig caecum and colon (range 617–748 sequences per host species). **a**, A neighbour-joining tree was created from 1,241 unambiguously aligned positions in all 2,735 sequences, with selected taxa indicated. Mollicutes, Lactobacillaceae and *Clostridium* clusters IV and XIVa are within the Firmicutes[96]. **b**, Host-specific trees were created with the same topology as the entire tree, shown in part **a**, but they depict only the sequences derived from the indicated host species. Branches shared with at least one other host species are shown in black, and branches specific to a single species are coloured. The same phyla and classes predominate in these animals (evident from the overlapping tree topologies and shared branches), although their relative abundances vary. By contrast, most genera and many families are specific to a single host species (coloured branches).

created a new niche for deadly pathogens. Ten of these 25 major infectious diseases could have arisen only after urbanization, because they depend on human–human transmission and quickly kill infected individuals or leave them with lifelong immunity[84]. Such 'crowd' diseases could not have survived in the small dispersed human societies present before agriculture. Common pathogens derived from human mutualistic microorganisms have also exploited these changes in the human population, with many clonal lineages being disseminated globally[13–18].

Urbanization and global travel have eroded some of the barriers to microbial transmission between social groups that contribute to the metacommunity structure of the human–microbe symbiosis. The diminished fidelity of host and symbiont lineages to each other (both within and between generations) and reduced opportunities for community-level selection between human social groups have reduced the strength of selection for mutualism. Microbial cheaters that allocate resources to their own growth and dissemination instead of pathogen interference or other costly contributions to host fitness can now spread globally, instead of merely within a tribe. Symbionts that colonize an infant who resides in an urban area include many microorganisms that are not derived from the infant's relatives, much less from an extended

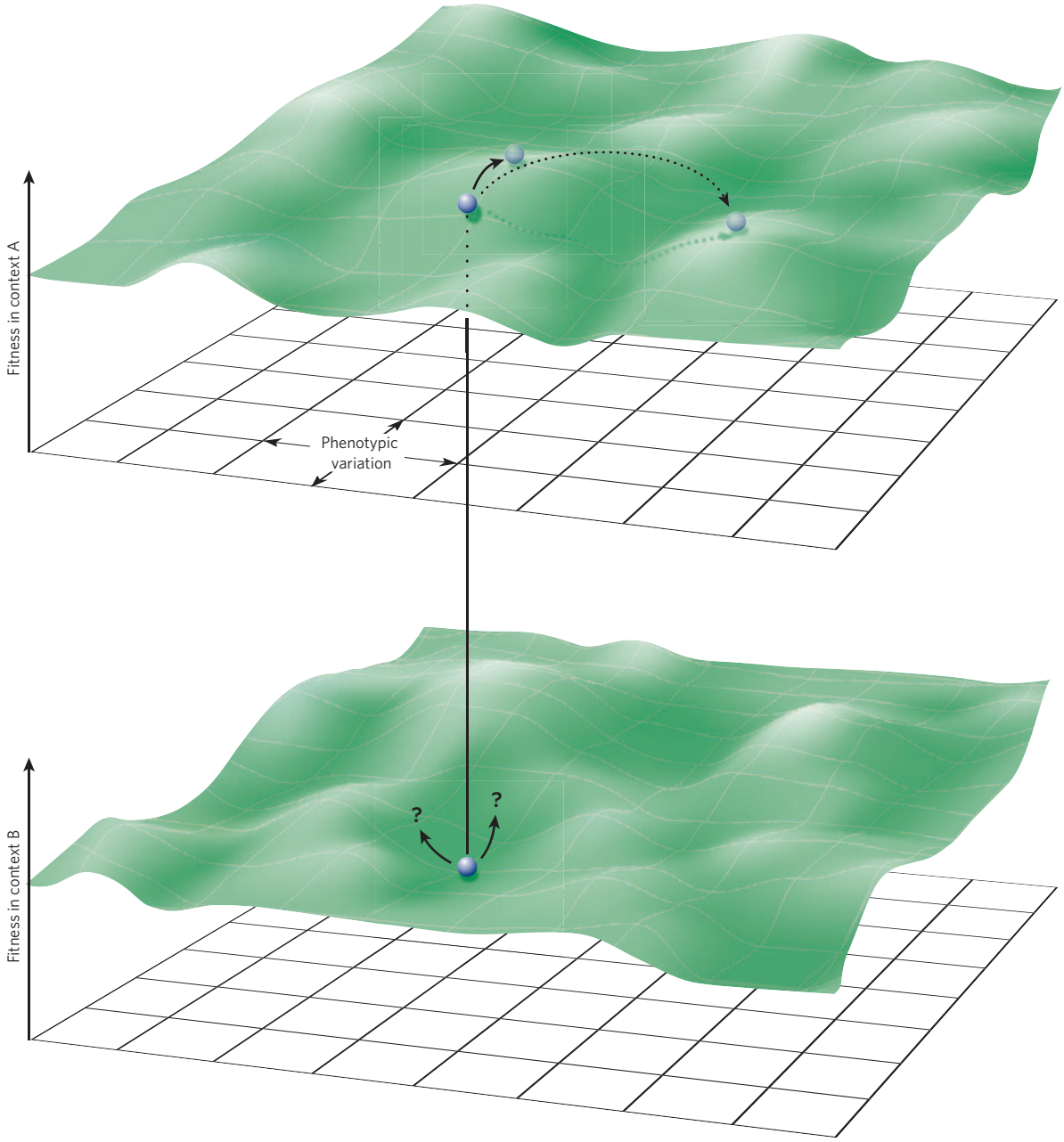


Figure 4 | Adaptive landscapes. The plane is a conceptual representation of the multidimensional phenotypes that are available to a microorganism. The height of the surface above the plane represents the fitness of the corresponding phenotypes in a given ecological context, including biotic and abiotic components of the environment. In a given environment (context A, upper panel), for mutations that have a small effect, a phenotype (circle) under natural selection will tend to evolve along the steepest path uphill towards higher fitness (solid arrow), eventually moving the mean phenotype of a population to a local fitness maximum. Mutations that have a large effect, such as horizontal gene transfer, can shift a phenotype to the slope of a different fitness peak (dashed arrow). This can markedly alter the outcome for the host; for example, it can result in pathogenesis instead of mutualism. The valley separating the peaks represents phenotypes of low fitness, such as those that are likely to elicit an immune response but lack the adaptations necessary to survive it. For a given phenotype, a change in context (for example, a change in host diet, alterations in coexisting microbial populations, or transfer to a different host or host species; context B, lower panel) can have subtle or marked effects on fitness. A phenotype near a fitness peak in context A might be in a valley of low fitness in context B. If the microorganism survives, the subsequent course of evolution might depend on the direction of phenotypic change caused by the next mutation.

kin group with a consistent lifestyle and geographic range over generations. This disruption of co-evolved mutualism between humans and human microbiota, as a result of changes in human ecology, contributes to the increasing prevalence of chronic and degenerative disease in industrialized countries[21,47].

Paths forward

Researchers have only just begun to describe the microbial communities that are associated with humans and the extent of the interactions between host and microbiota. Understanding this symbiotic 'landscape' will require research that spans the biological hierarchy from molecules to communities and is informed by ecological and evolutionary theory. Only with an integrated approach will it be possible to comprehend the complex ecology of human health and the many ways in which interactions between humans and microorganisms can go awry.

The first step in improving our understanding is to describe the composition of microbial communities in each habitat of the human body and how this varies over time, among individuals and with respect to variables such as diet, host genotype and health status. This project is now in its early stages, with the first successful forays having laid the groundwork for more ambitious studies, such as the Human Microbiome Project (see page 804).

Several recent studies highlight remarkable examples of how a co-evolved microbiota can markedly affect host biology at the molecular level[19–25], and these findings call for a complete re-examination of human physiology and immunology[44]. Attributes that were assumed to be human traits have been shown to result from human–microbe interactions.

Although human studies are essential, the technical and ethical limitations of carrying out experiments and obtaining samples from humans mean that experimental model systems also need to be used. These two approaches offer complementary information. The relevance of human studies is clear. But experimental model systems have two main advantages: they highlight evolutionarily conserved features that are likely to be crucial for function, and they show diversity (how a single 'goal' is accomplished differently), thereby exposing the essence of a characteristic. Models for the study of symbioses range from binary relationships between an invertebrate and one microbial species to complex vertebrate systems involving consortia of microorganisms (Table 1). For models with complex consortia, gnotobiotic techniques are used to manipulate the symbiosis experimentally. By contrast, using simpler consortia facilitates the molecular dissection of interactions in the intact natural setting. The genetic tools available for some model hosts allow the identification of genes and proteins that control host responses and manage the consortia.

From the microbial perspective, the host is a simply a complex environment — the distinction between human health and disease is important only as far as it affects microbial fitness. To think that we can intervene effectively in human–microbe relationships without considering microbial ecology and evolution is folly, as demonstrated by the spread of antibiotic-resistant microorganisms[13,14,16,17,58,59] and by the connections between some modern diseases and alterations in the human microbiota[21,47]. The principles and mechanisms that underlie microbial community structure and host–symbiont relationships must become incorporated into our definitions of human health. It will be crucial to consider the role of microbial communities, and not just individual species, as pathogens and mutualists[55]. Moreover, one of the goals of medical intervention during disease should be minimizing damage to the health-associated homeostasis between humans and their microbiota. Medical and general educational curricula will need to be modified accordingly. ■

1. Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I. & Dewhirst, F. E. Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* **43,** 5721–5732 (2005).
2. Bik, E. M. *et al.* Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl Acad. Sci. USA* **103,** 732–737 (2006).
3. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
4. Gao, Z., Tseng, C. H., Pei, Z. & Blaser, M. J. Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl Acad. Sci. USA* **104,** 2927–2932 (2007).
5. Pei, Z. *et al.* Bacterial biota in the human distal esophagus. *Proc. Natl Acad. Sci. USA* **101,** 4250–4255 (2004).
6. Verhelst, R. *et al.* Cloning of 16S rRNA genes amplified from normal and disturbed vaginal microflora suggests a strong association between *Atopobium vaginae*, *Gardnerella vaginalis* and bacterial vaginosis. *BMC Microbiol.* **4,** 16 (2004).
7. Zhou, X. *et al.* Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. *Microbiology* **150,** 2565–2573 (2004).
8. Lay, C. *et al.* Colonic microbiota signatures across five northern European countries. *Appl. Environ. Microbiol.* **71,** 4153–4155 (2005).
9. Matsuki, T., Watanabe, K., Fujimoto, J., Takada, T. & Tanaka, R. Use of 16S rRNA gene-targeted group-specific primers for real-time PCR analysis of predominant bacteria in human feces. *Appl. Environ. Microbiol.* **70,** 7220–7228 (2004).
10. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5,** e177 (2007).
11. Vanhoutte, T., Huys, G., De Brandt, E. & Swings, J. Temporal stability analysis of the microbiota in human feces by denaturing gradient gel electrophoresis using universal and group-specific 16S rRNA gene primers. *FEMS Microbiol. Ecol.* **48,** 437–446 (2004).
12. Zoetendal, E. G., Akkermans, A. D. L., Akkermans-van Vliet, W. M., de Visser, J. A. G. M. & de Vos, W. M. The host genotype affects the bacterial community in the human gastrointestinal tract. *Microb. Ecol. Health Dis.* **13,** 129–134 (2001).
13. Leavis, H. L., Bonten, M. J. & Willems, R. J. Identification of high-risk enterococcal clonal complexes: global dispersion and antibiotic resistance. *Curr. Opin. Microbiol.* **9,** 454–460 (2006).
14. Miragaia, M., Thomas, J. C., Couto, I., Enright, M. C. & de Lencastre, H. Inferring a population structure for *Staphylococcus epidermidis* from multilocus sequence typing data. *J. Bacteriol.* **189,** 2540–2552 (2007).
15. Callaghan, M. J., Jolley, K. A. & Maiden, M. C. Opacity-associated adhesin repertoire in hyperinvasive *Neisseria meningitidis*. *Infect. Immun.* **74,** 5085–5094 (2006).
16. Robinson, D. A. & Enright, M. C. Multilocus sequence typing and the evolution of methicillin-resistant *Staphylococcus aureus*. *Clin. Microbiol. Infect.* **10,** 92–97 (2004).
17. Robinson, D. A., Sutcliffe, J. A., Tewodros, W., Manoharan, A. & Bessen, D. E. Evolution and global dissemination of macrolide-resistant group A streptococci. *Antimicrob. Agents Chemother.* **50,** 2903–2911 (2006).
18. Wirth, T. *et al.* Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol. Microbiol.* **60,** 1136–1151 (2006).
19. Bäckhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl Acad. Sci. USA* **101,** 15718–15723 (2004).
20. Cash, H. L., Whitham, C. V., Behrendt, C. L. & Hooper, L. V. Symbiotic bacteria direct expression of an intestinal bactericidal lectin. *Science* **313,** 1126–1130 (2006).
21. Guarner, F. *et al.* Mechanisms of disease: the hygiene hypothesis revisited. *Nature Clin. Pract. Gastroenterol. Hepatol.* **3,** 275–284 (2006).
22. Kelly, D. *et al.* Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear–cytoplasmic shuttling of PPAR-γ and RelA. *Nature Immunol.* **5,** 104–112 (2004).
23. Martin, F. P. *et al.* A top-down systems biology view of microbiome–mammalian metabolic interactions in a mouse model. *Mol. Syst. Biol.* **3,** 112 (2007).
24. Mazmanian, S. K., Liu, C. H., Tzianabos, A. O. & Kasper, D. L. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* **122,** 107–118 (2005).
25. Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. & Medzhitov, R. Recognition of commensal microflora by Toll-like receptors is required for intestinal homeostasis. *Cell* **118,** 229–241 (2004).
26. Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124,** 837–848 (2006).
27. Gong, J. *et al.* 16S rRNA gene-based analysis of mucosa-associated bacterial community and phylogeny in the chicken gastrointestinal tracts: from crops to ceca. *FEMS Microbiol. Ecol.* **59,** 147–157 (2007).
28. Mackie, R. I., Rycyk, M., Ruemmler, R. L., Aminov, R. I. & Wikelski, M. Biochemical and microbiological evidence for fermentative digestion in free-living land iguanas (*Conolophus pallidus*) and marine iguanas (*Amblyrhynchus cristatus*) on the Galapagos archipelago. *Physiol. Biochem. Zool.* **77,** 127–138 (2004).
29. Nelson, K. E. *et al.* Phylogenetic analysis of the microbial populations in the wild herbivore gastrointestinal tract: insights into an unexplored niche. *Environ. Microbiol.* **5,** 1212–1220 (2003).
30. Uenishi, G. *et al.* Molecular analyses of the intestinal microbiota of chimpanzees in the wild and in captivity. *Am. J. Primatol.* **69,** 367–376 (2007).
31. Wilson, K. H., Brown, R. S., Andersen, G. L., Tsang, J. & Sartor, B. Comparison of fecal biota from specific pathogen free and feral mice. *Anaerobe* **12,** 249–253 (2006).
32. Wilson, D. S. Biological communities as functionally organized units. *Ecology* **78,** 2018–2024 (1997).
33. Foster, K. R. & Wenseleers, T. A general model for the evolution of mutualisms. *J. Evol. Biol.* **19,** 1283–1293 (2006).
34. Sachs, J. L., Mueller, U. G., Wilcox, T. P. & Bull, J. J. The evolution of cooperation. *Q. Rev. Biol.* **79,** 135–160 (2004).
35. Flint, H. J. Polysaccharide breakdown by anaerobic microorganisms inhabiting the mammalian gut. *Adv. Appl. Microbiol.* **56,** 89–120 (2004).
36. Flint, H. J., Duncan, S. H., Scott, K. P. & Louis, P. Interactions and competition within the microbial community of the human colon: links between diet and health. *Environ. Microbiol.* **9,** 1101-1111 (2007).
37. Fons, M., Gomez, A. & Karjalainen, T. Mechanisms of colonisation and colonisation resistance of the digestive tract. Part 2: bacteria/bacteria interactions. *Microb. Ecol. Health Dis.* **12,** 240–246 (2000).
38. Reid, G. & Bruce, A. W. Probiotics to prevent urinary tract infections: the rationale and evidence. *World J. Urol.* **24,** 28–32 (2006).
39. Brook, I. The role of bacterial interference in otitis, sinusitis and tonsillitis. *Otolaryngol. Head Neck Surg.* **133,** 139–146 (2005).
40. Servin, A. L. Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. *FEMS Microbiol. Rev.* **28,** 405–440 (2004).

41. Tilman, D. Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proc. Natl Acad. Sci. USA* **101,** 10854–10861 (2004).
42. Pool-Zobel, B., Veeriah, S. & Bohmer, F. D. Modulation of xenobiotic metabolising enzymes by anticarcinogens — focus on glutathione *S*-transferases and their role as targets of dietary chemoprevention in colorectal carcinogenesis. *Mutat. Res.* **591,** 74–92 (2005).
43. Doebeli, M., Hauert, C. & Killingback, T. The evolutionary origin of cooperators and defectors. *Science* **306,** 859–862 (2004).
44. McFall-Ngai, M. Adaptive immunity: care for the community. *Nature* **445,** 153 (2007).
45. Macpherson, A. J., Geuking, M. B. & McCoy, K. D. Immune responses that adapt the intestinal mucosa to commensal intestinal bacteria. *Immunology* **115,** 153–162 (2005).
46. Matzinger, P. The danger model: a renewed sense of self. *Science* **296,** 301–305 (2002).
47. O'Keefe, S. J. *et al.* Why do African Americans get more colon cancer than Native Africans? *J. Nutr.* **137,** 175S–182S (2007).
48. Moore, W. E. & Moore, L. H. Intestinal floras of populations that have a high risk of colon cancer. *Appl. Environ. Microbiol.* **61,** 3202–3207 (1995).
49. Swenson, W., Wilson, D. S. & Elias, R. Artificial ecosystem selection. *Proc. Natl Acad. Sci. USA* **97,** 9110–9114 (2000).
50. Dethlefsen, L., Eckburg, P. B., Bik, E. M. & Relman, D. A. Assembly of the human intestinal microbiota. *Trends Ecol. Evol.* **21,** 517–523 (2006).
51. Young, V. B. & Schmidt, T. M. Antibiotic-associated diarrhea accompanied by large-scale alterations in the composition of the fecal microbiota. *J. Clin. Microbiol.* **42,** 1203–1206 (2004).
52. Li, J. *et al.* Identification of early microbial colonizers in human dental biofilm. *J. Appl. Microbiol.* **97,** 1311–1318 (2004).
53. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
54. Klaassens, E. S., de Vos, W. M. & Vaughan, E. E. Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Appl. Environ. Microbiol.* **73,** 1388–1392 (2007).
55. Lepp, P. W. *et al.* Methanogenic Archaea and human periodontal disease. *Proc. Natl Acad. Sci. USA* **101,** 6176–6181 (2004).
56. Jernberg, C., Sullivan, A., Edlund, C. & Jansson, J. K. Monitoring of antibiotic-induced alterations in the human intestinal microflora and detection of probiotic strains by use of terminal restriction fragment length polymorphism. *Appl. Environ. Microbiol.* **71,** 501–506 (2005).
57. Pepin, J. *et al.* Emergence of fluoroquinolones as the predominant risk factor for *Clostridium difficile*-associated diarrhea: a cohort study during an epidemic in Quebec. *Clin. Infect. Dis.* **41,** 1254–1260 (2005).
58. Lofmark, S., Jernberg, C., Jansson, J. K. & Edlund, C. Clindamycin-induced enrichment and long-term persistence of resistant *Bacteroides* spp. and resistance genes. *J. Antimicrob. Chemother.* **58,** 1160–1167 (2006).
59. Sjolund, M., Tano, E., Blaser, M. J., Andersson, D. I. & Engstrand, L. Persistence of resistant *Staphylococcus epidermidis* after single course of clarithromycin. *Emerg. Infect. Dis.* **11,** 1389–1393 (2005).
60. Kolenbrander, P. E. *et al.* Bacterial interactions and successions during plaque development. *Periodontol. 2000* **42,** 47–79 (2006).
61. Savage, D. C. in *Mucosal Immunology* (eds Mestecky, J. *et al.*) 19-34 (Elsevier, Boston, 2005).
62. Caufield, P. W. *et al.* Natural history of *Streptococcus sanguinis* in the oral cavity of infants: evidence for a discrete window of infectivity. *Infect. Immun.* **68,** 4018–4023 (2000).
63. Samuel, B. S. & Gordon, J. I. A humanized gnotobiotic mouse model of host–Archaeal-bacterial mutualism. *Proc. Natl Acad. Sci. USA* **103,** 10011–10016 (2006).
64. Kolenbrander, P. E. *et al.* Communication among oral bacteria. *Microbiol. Mol. Biol. Rev.* **66,** 486–505 (2002).
65. Xu, J. *et al.* A genomic view of the human–*Bacteroides thetaiotaomicron* symbiosis. *Science* **299,** 2074–2076 (2003).
66. Schell, M. A. *et al.* The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract. *Proc. Natl Acad. Sci. USA* **99,** 14422–14427 (2002).
67. Czárán, T. L., Hoekstra, R. F. & Pagie, L. Chemical warfare between microbes promotes biodiversity. *Proc. Natl Acad. Sci. USA* **99,** 786–790 (2002).
68. Gordon, D. M., Riley, M. A. & Pinou, T. Temporal changes in the frequency of colicinogeny in *Escherichia coli* from house mice. *Microbiology* **144,** 2233–2240 (1998).
69. Sperandio, V., Torres, A. G., Jarvis, B., Nataro, J. P. & Kaper, J. B. Bacteria-host communication: the language of hormones. *Proc. Natl Acad. Sci. USA* **100,** 8951–8956 (2003).
70. Shiner, E. K., Rumbaugh, K. P. & Williams, S. C. Inter-kingdom signaling: deciphering the language of acyl homoserine lactones. *FEMS Microbiol. Rev.* **29,** 935–947 (2005).
71. Rendon, M. A. *et al.* Commensal and pathogenic *Escherichia coli* use a common pilus adherence factor for epithelial cell colonization. *Proc. Natl Acad. Sci. USA* **104,** 10637–10642 (2007).
72. Wren, B. W. The yersiniae — a model genus to study the rapid evolution of bacterial pathogens. *Nature Rev. Microbiol.* **1,** 55–64 (2003).
73. Monot, M. *et al.* On the origin of leprosy. *Science* **308,** 1040–1042 (2005).
74. Brown, N. F., Wickham, M. E., Coombes, B. K. & Finlay, B. B. Crossing the line: selection and evolution of virulence traits. *PLoS Pathog.* **2,** e42 (2006).
75. Woolhouse, M. E., Webster, J. P., Domingo, E., Charlesworth, B. & Levin, B. R. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genet.* **32,** 569–577 (2002).
76. Wickham, M. E., Brown, N. F., Boyle, E. C., Coombes, B. K. & Finlay, B. B. Virulence is positively selected by transmission success between mammalian hosts. *Curr. Biol.* **17,** 783–788 (2007).
77. Walther, B. A. & Ewald, P. W. Pathogen survival in the external environment and the evolution of virulence. *Biol. Rev. Camb. Philos. Soc.* **79,** 849–869 (2004).
78. Boots, M. & Mealor, M. Local interactions select for lower pathogen infectivity. *Science* **315,** 1284–1286 (2007).
79. Taylor, L. H., Latham, S. M. & Woolhouse, M. E. Risk factors for human disease emergence. *Phil. Trans. R. Soc. Lond. B* **356,** 983–989 (2001).
80. Naylor, S. W., Gally, D. L. & Low, J. C. Enterohaemorrhagic *E. coli* in veterinary medicine. *Int. J. Med. Microbiol.* **295,** 419–441 (2005).
81. Read, A. F. & Taylor, L. H. The ecology of genetically diverse infections. *Science* **292,** 1099–1102 (2001).
82. West, S. A. & Buckling, A. Cooperation, virulence and siderophore production in bacterial parasites. *Proc. R. Soc. Lond B* **270,** 37–44 (2003).
83. Gardner, A., West, S. A. & Buckling, A. Bacteriocins, spite and virulence. *Proc. R. Soc. Lond B* **271,** 1529–1535 (2004).
84. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447,** 279–283 (2007).
85. Woolhouse, M. E., Taylor, L. H. & Haydon, D. T. Population biology of multihost pathogens. *Science* **292,** 1109–1112 (2001).
86. Cheesman, S. E. & Guillemin, K. We know you are in there: conversing with the indigenous gut microbiota. *Res. Microbiol.* **158,** 2–9 (2007).
87. Hongoh, Y. *et al.* Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host. *Appl. Environ. Microbiol.* **71,** 6590–6599 (2005).
88. Kikuchi, Y. & Graf, J. Spatial and temporal population dynamics of a naturally occurring two-species microbial community inside the digestive tract of the medicinal leech. *Appl. Environ. Microbiol.* **73,** 1984–1991 (2007).
89. Broderick, N. A., Raffa, K. F. & Handelsman, J. Midgut bacteria required for *Bacillus thuringiensis* insecticidal activity. *Proc. Natl Acad. Sci. USA* **103,** 15196–15199 (2006).
90. Cox, C. R. & Gilmore, M. S. Native microbial colonization of *Drosophila melanogaster* and its use as a model of *Enterococcus faecalis* pathogenesis. *Infect. Immun.* **75,** 1565–1576 (2007).
91. Fraune, I. & Bosch, T. Long-term maintenance of species-specific bacterial microbiota in the basal metazoan Hydra. *Proc. Natl Acad. Sci. USA* **104,** 13146–13151
92. Nyholm, S. V. & McFall-Ngai, M. J. The winnowing: establishing the squid–*Vibrio* symbiosis. *Nature Rev. Microbiol.* **2,** 632–642 (2004).
93. Davidson, S. K. & Stahl, D. A. Transmission of nephridial bacteria of the earthworm *Eisenia fetida*. *Appl. Environ Microbiol.* **72,** 769–775 (2006).
94. Goodrich-Blair, H. & Clarke, D. J. Mutualism and pathogenesis in *Xenorhabdus* and *Photorhabdus*: two roads to the same destination. *Mol. Microbiol.* **64,** 260–268 (2007).
95. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).
96. Collins, M. D. *et al.* The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *Int. J. Syst. Bacteriol.* **44,** 812–826 (1994).

Acknowledgements Research in the laboratory of D.A.R. is supported by funds from the Doris Duke Charitable Foundation, the Horn Foundation, the Office of Naval Research and the National Institutes of Health (NIH). Research in the laboratory of M.M.-N. is supported by the NIH and the National Science Foundation. D.A.R. is a recipient of an NIH Director's Pioneer Award and a Doris Duke Distinguished Clinical Scientist Award.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to D.A.R. (relman@stanford.edu).

Recognition of microorganisms and activation of the immune response

Ruslan Medzhitov[1]

The mammalian immune system has innate and adaptive components, which cooperate to protect the host against microbial infections. The innate immune system consists of functionally distinct 'modules' that evolved to provide different forms of protection against pathogens. It senses pathogens through pattern-recognition receptors, which trigger the activation of antimicrobial defences and stimulate the adaptive immune response. The adaptive immune system, in turn, activates innate effector mechanisms in an antigen-specific manner. The connections between the various immune components are not fully understood, but recent progress brings us closer to an integrated view of the immune system and its function in host defence.

Infectious diseases are a leading cause of morbidity and mortality worldwide and are a major challenge for the biomedical sciences. Improved sanitary conditions, clean water supplies and vector control are by far the most effective measures to reduce the incidence of infectious disease. However, the development of vaccines and therapeutics is also important, and this requires an understanding of the host immune system. Recently, much progress has been made towards discovering the mechanisms of microbial pathogenesis and host–microbe symbiosis. And knowledge about the immune system has also been steadily increasing. Yet many challenges remain, perhaps the most daunting being effective vaccine development. Indeed, it is not known how to elicit protective immunity against most pathogens in a safe and practical manner. To achieve this and other goals, such as the safe and efficient blockade of autoimmune and allergic immune responses, further developments in basic research are clearly required.

Here, I provide a general overview of the immune system as it relates to defence against microorganisms, with an emphasis on recent findings.

Host–microbe interactions

All metazoan hosts exist in close association with microbial communities that colonize them. The 'rules of engagement' of host–microbe interactions are incompletely understood, and defining these is clearly important for understanding the evolution and functioning of the immune system.

The host as a set of niches colonized by microorganisms

Mammalian hosts provide a number of niches that can be colonized by microorganisms, including the skin, intestine, upper and lower respiratory tract, urogenital tract and internal organs. Some of these niches (for example, the colon and the skin) are colonized constitutively by an endogenous microbiota. Other niches (for example, the internal organs and the lower respiratory tract) are normally kept sterile (in an immunocompetent host). The effect of microbial colonization on host fitness depends on the microbial adaptation strategy. These effects can be positive, as is the case for the many intestinal bacteria that provide a range of benefits to the host (see pages 804 and 811). In other cases, microbial colonization can be detrimental to the host, and these colonizing bacteria are referred to as pathogens. Such negative effects can depend on the status of the host's immune system: for example, certain pathogens, known as opportunistic pathogens, affect only immunocompromised individuals.

Virulence factors

The adaptation of bacteria to particular host niches depends on the activity of various adaptation factors; for pathogens, these are known as virulence factors. Adaptation factors are often encoded on mobile genetic elements (for example, plasmids and genomic islands) that can be transmitted within and between bacterial species[1] (see page 835), although there are important exceptions (for example, in *Mycobacterium* spp.)[2]. The role of virulence factors is to enable adaptation to the specific environments in the host niches and to promote transmission to another host. In this way, some common themes of virulence-factor activity (and therefore pathogenicity) can be identified[3]. Depending on the niche that they colonize, bacterial pathogens have virulence factors that allow a range of activities: penetration of surface epithelia, attachment to cell surfaces and/or the extracellular matrix, invasion of intracellular compartments, acquisition of iron, evasion of host-defence mechanisms and transmission to another host. Different strategies of pathogenic microbial adaptation are associated with varying degrees of damage to the tissues of the host. Regardless of the degree of virulence, at least some symptoms of infectious disease are side-effects of microbial adaptation to host niches.

Recognition of microorganisms by the immune system

The detrimental effects of microbial infections led to the evolution of a variety of host-defence mechanisms. In jawed vertebrates, there are two types of defence: innate and adaptive (also known as acquired). The main distinction between these is the receptor types used to recognize pathogens. Innate immune recognition is mediated by pattern-recognition receptors (PRRs), which are germline encoded, and each receptor has broad specificities for conserved and invariant features of microorganisms[4]. By contrast, adaptive immune recognition is mediated by antigen receptors: the genes encoding these receptors are assembled from gene segments in the germ line, and somatic recombination of these segments enables the generation of a diverse repertoire of receptors with random but narrow specificities[5]. Antigen receptors are clonally distributed on T and B lymphocytes, which allows clonal selection of pathogen-specific receptors and is the basis for immunological memory. (That is, each lymphocyte expresses antigen receptors of a single specificity, so only specific populations of lymphocytes are selected to expand in response to a pathogen.) Therefore, the innate immune system and the adaptive immune system deal with the molecular diversity of pathogens in fundamentally different ways.

[1]Howard Hughes Medical Institute and Department of Immunobiology, Yale University School of Medicine, TAC S-669, 300 Cedar Street, New Haven, Connecticut 06510, USA.

Innate immune system

Innate immune recognition (also known as pattern recognition) is based on the detection of molecular structures that are unique to microorganisms[4]. Pattern recognition is unusual in that each host receptor (PRR) has a broad specificity and can potentially bind to a large number of molecules that have a common structural motif or pattern. The targets of PRRs are sometimes referred to as pathogen-associated molecular patterns (PAMPs), although they are present on both pathogenic and non-pathogenic microorganisms. PAMPs are well suited to innate immune recognition for three main reasons. First, they are invariant among microorganisms of a given class. Second, they are products of pathways that are unique to microorganisms, allowing discrimination between self and non-self molecules. Third, they have essential roles in microbial physiology, limiting the ability of the microorganisms to evade innate immune recognition through adaptive evolution of these molecules. Bacterial PAMPs are often components of the cell wall, such as lipopolysaccharide, peptidoglycan, lipoteichoic acids and cell-wall lipoproteins. An important fungal PAMP is β-glucan, which is a component of fungal cell walls. The detection of these structures by the innate immune system can signal the presence of microorganisms. The recognition of viruses also partly follows this principle. However, because all viral components are synthesized within host cells, the main targets of innate immune recognition in this case are viral nucleic acids. Discrimination between self (host) and viral nucleic acids occurs on the basis of specific chemical modifications and structural features that are unique to viral RNA and DNA, as well as on the cellular compartments where viral (but not host-derived) nucleic acids are normally found (discussed later). Nevertheless, this discrimination is not perfect and can fail under certain conditions, which can result in the development of autoimmune diseases[6].

An important aspect of pattern recognition is that PRRs themselves do not distinguish between pathogenic microorganisms and symbiotic (non-pathogenic) microorganisms, because the ligands of the receptors are not unique to pathogens. Yet, despite humans being colonized by trillions of symbiotic bacteria, homeostasis is somehow maintained under normal conditions. Furthermore, innate immune recognition of symbiotic microorganisms has an important role in maintaining intestinal homeostasis[7]. And dysregulation of these interactions can lead to the development of inflammatory bowel disease and other disorders.

PRRs and their functions

There are several functionally distinct classes of PRR (Table 1). The best characterized class is Toll-like receptors (TLRs). TLRs are transmembrane receptors that recognize viral nucleic acids and several bacterial products, including lipopolysaccharide and lipoteichoic acids (see ref. 8 for a review). The full range of TLR functions in antimicrobial defence has not yet been determined, but TLRs are known to elicit inflammatory and antimicrobial responses after activation by their microbial ligands.

In terms of the inflammatory response, TLRs activate tissue-resident macrophages to produce pro-inflammatory cytokines, including tumour-necrosis factor (TNF), interleukin-1β (IL-1β) and IL-6, which coordinate local and systemic inflammatory responses. TNF and IL-1β, in turn, activate the local endothelium to induce vasodilation and increase the permeability of the blood vessel, allowing serum proteins and leukocytes to be recruited to the site of infection. In addition, an increase in the amount of tissue factor (also known as coagulation factor III) on the endothelium leads to a local coagulation cascade that helps to prevent microbial dissemination through the blood. Furthermore, IL-1β, together with IL-6, activates hepatocytes to produce acute-phase proteins, including collectins and pentraxins. These proteins, in turn, activate complement and opsonize pathogens for phagocytosis by macrophages and neutrophils. In this way, TLRs indirectly elicit an antimicrobial response.

TLRs also directly trigger such a response, by inducing macrophages to produce antimicrobial proteins and peptides. In mouse macrophages, the activation of TLRs results in transcription of the gene encoding inducible nitric-oxide synthase (iNOS; also known as NOS2), which has an important role in antimicrobial defence[9]. Interestingly, iNOS is not produced in response to the activation of TLRs on human macrophages[10]. Instead, human keratinocytes synthesize vitamin D, which is crucial for antimicrobial activity, partly because of vitamin-D-receptor-dependent induction of the gene encoding the antimicrobial peptide LL37 (also known as CAMP)[11]. Sunlight (a source of UVB radiation) is necessary for vitamin D synthesis, so the difference in vitamin D requirement for antimicrobial defence might reflect the nocturnal and diurnal lifestyles of mice and humans, respectively.

Another well-characterized PRR is dectin 1, a transmembrane receptor that binds to β-glucan[12] and is present on dendritic cells and macrophages. Dectin 1 is a member of a large family of C-type lectins, many of which are present on these same cell types but have unknown functions[13]. Dectin 1 contains an atypical immunoreceptor tyrosine-based activation motif (ITAM) that engages the protein tyrosine kinase SYK, thereby activating a signalling pathway that involves CARD9, Bcl-10 and MALT1 (ref. 14). This PRR has an important role in antifungal defence[15,16], being involved in the phagocytosis of fungal pathogens, the induction of an antimicrobial response (such as activation of NADPH oxidase) and the production of cytokines[17].

In addition to transmembrane receptors on the cell surface and in endosomal compartments, there are intracellular (cytosolic) receptors that function in the pattern recognition of bacterial and viral pathogens. These include NLRs and the intracellular sensors of viral nucleic acids RIG-I (retinoic-acid-inducible gene I; also known as DDX58), MDA5 (melanoma differentiation-associated gene 5; also known as IFIH1) and DAI (DNA-dependent activator of interferon-regulatory factors; also known as ZBP1). NLRs are a large family of about 20 intracellular proteins with a common protein-domain organization but diverse functions[18–21]. All NLRs contain a nucleotide-binding oligomerization domain (NOD) followed by a leucine-rich-repeat domain at the carboxy terminus. At the amino terminus, NLRs have one of three domains and are thereby categorized into three subfamilies: a caspase-recruitment domain (CARD), present in proteins in the NOD subfamily; a pyrin domain, in the NALP subfamily; or a BIR domain (baculoviral inhibitor-of-apoptosis-protein repeat-containing domain), in the NAIP subfamily[18–21]. The N-terminal domains engage distinct signalling pathways, which define the functional properties of the family members.

The proteins of the NOD subfamily — NOD1 and NOD2 — are both involved in sensing bacterial peptidoglycans, although they recognize structurally distinct peptidoglycan fragments[18]. The sensing of peptidoglycan by NOD1 or NOD2 triggers the production of pro-inflammatory cytokines and chemokines and the recruitment of neutrophils to the site of infection[19]. In addition, these NOD proteins contribute to the initiation of the adaptive immune response[22,23], and mutations in NOD2 have been implicated in the pathogenesis of Crohn's disease[24]. NOD2 is also crucial for the production of antimicrobial peptides known as defensins by Paneth cells (which are present in the small intestine), and NOD proteins can presumably activate antimicrobial responses in other cell types[23].

The NALP subfamily of NLRs has 14 members, and at least some of these are involved in the induction of the inflammatory response mediated by the IL-1 family of cytokines, which includes IL-1β, IL-18 and IL-33 (ref. 20). These cytokines are synthesized as inactive precursors that need to be cleaved by the pro-inflammatory caspases: that is, caspase 1, caspase 4 and caspase 5 in humans, and caspase 1, caspase 11 and caspase 12 in mice. These caspases are activated in a multisubunit complex called the inflammasome[25]. There are several types of inflammasome, categorized according to their composition and the involvement of a particular NALP or NAIP. The individual inflammasomes are activated in response to a variety of bacterial infections, by mechanisms that have been poorly defined[20]. Why IL-1-family members are activated by such an elaborate mechanism is puzzling. Unlike other pro-inflammatory cytokines, IL-1β production is regulated by two distinct signals: TLR-induced transcription and inflammasome-mediated processing of the precursor protein. It is possible that, in addition to IL-1-family members, the inflammasomes

Table 1 | Modules of the innate immune system

Innate host-defence module	Primary sensors (PRRs)	Prototypical responses
Mucosal epithelia	TLRs and NOD proteins	Production of antimicrobial peptides Production of mucins
Phagocytes	TLRs, dectins and NOD proteins	Production of antimicrobial proteins Production of cytokines: IL-1β, IL-6 and TNF
Acute-phase proteins and complement system	Collectins, pentraxins and ficolins	Lysis or opsonization of pathogens Chemotactic attraction of leukocytes
Inflammasomes	NALPs and NAIPs	Production of IL-1-family members Apoptosis of infected host cells
NK cells	ND	Apoptosis of infected host cells
Type-I-IFN-induced antiviral proteins	RIG-I, MDA5, DAI and TLRs	Induction of an antiviral state Apoptosis of infected host cells
Eosinophils and basophils	ND	Contraction of smooth muscle Production of mucins Peristalsis Production of biogenic amines Production of cytokines: IL-4, IL-5, IL-9, IL-13 and TNF
Mast cells	ND	Contraction of smooth muscle Production of mucins Peristalsis Production of biogenic amines Production of cytokines: IL-4, IL-5, IL-9, IL-13 and TNF

This list of modules, sensors and responses is not comprehensive and has been simplified for clarity. It should be noted that the function of NALPs and NAIPs is not completely understood. In addition, the primary sensors that control expression of NK-cell-receptor ligands, as well as the sensors that activate antihelminthic responses by mast cells, eosinophils and basophils, have not been identified. Certain modules can be co-induced during infection; these modules are functionally linked (for example, phagocytes and the complement system) and co-regulated by the same cytokines. ND, not determined.

process antimicrobial peptides or proteins that have not yet been characterized. NALPs might also contribute to antimicrobial defence by inducing the apoptosis of infected cells[20]. Whether they can also directly induce the expression of antimicrobial genes is unknown.

Intracellular recognition of viral infections is mediated by two types of viral nucleic-acid sensor. Viral RNA in the cytosol is detected by the RNA-helicase-family proteins RIG-I and MDA5 (ref. 26), whereas viral DNA is detected by the recently identified protein DAI[27]. RIG-I and MDA5 recognize different types of viral RNA: single-stranded RNA containing 5′ triphosphate and double-stranded RNA, respectively[28–31]. These structural features are absent from cellular (host) RNAs, which contain either short hairpin structures, in the case of transfer RNAs and ribosomal RNAs, or a 5′-cap structure, in the case of messenger RNA. These structural differences allow discrimination between viral and self RNAs. Activation of RIG-I or MDA5 results in the production of type I interferons (IFNs; IFN-α and IFN-β) and thereby the induction of antiviral immunity[30]. Interestingly, a crucial adaptor involved in RIG-I and MDA5 signalling is associated with the mitochondrial membrane[32], but the reason for this is unclear at present. The details of how viral DNA is recognized in the cytosol, and the signalling pathways induced by the engagement of DAI, are not yet known. It is, however, clear that the RNA-sensing pathway and the DNA-sensing pathway converge on the protein kinase TBK1 (TANK-binding kinase 1) and the transcription factor IFN-regulatory factor 3 (refs 33–35). Type I IFNs are therefore elicited by the engagement of either type of sensor. This results in antiviral immune responses in both cases, through inducing the expression of numerous IFN-inducible genes, the products of which have a broad range of antiviral activities[36].

Adaptive immune system

Adaptive immune recognition is mediated by two types of antigen receptor: T-cell receptors and B-cell receptors. The genes encoding antigen receptors are assembled from variable and constant fragments through recombination-activating gene (RAG)-protein-mediated somatic recombination[5], a process that yields a diverse repertoire of receptors. This diversity is further increased by additional mechanisms, such as non-templated nucleotide addition, gene conversion and (in the case of B cells) somatic hypermutation, generating a highly diverse repertoire of receptors with the potential to recognize almost any antigenic determinant in a specific manner[5].

There are two types of lymphocyte that express antigen receptors: conventional lymphocytes and innate-like lymphocytes. In the case of conventional lymphocytes — that is, conventional T cells (most αβ T cells) and B cells (also known as B2 cells) — antigen receptors are assembled essentially at random. By contrast, for innate-like lymphocytes — that is, B1 cells, marginal-zone B cells, natural-killer T cells and subsets of γδ T cells — the diversity of antigen receptors is restricted and not entirely random. Their specificities are skewed towards a predefined set of ligands[37].

The specificities of the receptors of conventional lymphocytes are not predetermined and neither, therefore, is the site where these cells might encounter their cognate antigen (that is, the antigen specifically recognized by the receptor) or the effector response they need to elicit on activation. So these lymphocytes circulate through the lymph nodes, which drain most of the body's tissues and organs, and the spleen, which filters the blood, until they encounter an antigen that they are specific for. Microbial antigens are taken up by antigen-presenting cells in the peripheral tissues and are delivered to the lymph nodes or spleen through the lymph or blood, respectively, where they are recognized by conventional lymphocytes. Because the specificity of each antigen receptor is not directly linked to the origin of the antigen, conventional lymphocytes need to be able to differentiate into several types of effector cell, depending on the class of pathogen they recognize (discussed later). The differentiation of conventional lymphocytes into a particular effector-cell type and their localization to the site of infection are regulated by the instructions provided by the innate immune system, generally in the form of cytokines and chemokines, respectively.

There are two types of conventional αβ T cell: T-helper (T_H) cells, which are marked by the co-receptor CD4 on the cell surface; and cytotoxic T cells, which express CD8. These cells recognize antigenic peptides bound to major histocompatibility complex (MHC) class II and class I molecules, respectively. Conventional B cells can recognize almost any antigen by binding to a specific three-dimensional molecular determinant (or epitope).

Innate-like lymphocytes differ from conventional lymphocytes in several important ways. Although the antigen receptors of innate-like lymphocytes are assembled in a similar manner to those of conventional lymphocytes, their assembly process is not entirely random. Receptor diversity is biased towards a characteristic set of specificities for each subset of innate-like lymphocytes[37]. Accordingly, the effector functions of these lymphocytes and the sites where they reside are often predetermined. The effector responses of innate-like lymphocytes therefore do not generally require the same types of instruction that are provided by the innate immune system to conventional lymphocytes.

The innate-like B cells known as B1 cells reside in the peritoneal and pleural cavities and produce mainly antibodies of the IgM class with specificities skewed towards some common bacterial polysaccharides and some self antigens[38]. Innate-like T cells recognize non-classical MHC molecules (also known as MHC class Ib molecules), which can present bacteria-specific ligands: for example, bacterial lipids or formylated peptides in the case of the CD1 and H–2M3 families, respectively. In a way, these MHC-like molecules function as PRRs, presenting microbial ligands to specialized T cells[39]. Some non-classical MHC molecules might themselves be ligands for T-cell receptors, without presenting any other molecules. In this case, the expression of these molecules is thought to be inducible by the engagement of PRRs on specific cell types, such as mucosal epithelial cells[40].

Modules of the innate immune system

Unlike the adaptive immune system, the innate immune system is not a single entity. It is a collection of distinct subsystems, or modules, that appeared at different stages of evolution and carry out different functions in host defence. Some of the main modules found in mammals and how these function in innate host defence are described in this section (Table 1).

Mucosal epithelia

All metazoans have mucosal epithelia, one of the most ancient and universal modules of innate immunity. Together with the skin, the mucosal epithelia are the main interface between the host and the microbial world (including both pathogenic and symbiotic microorganisms). Mucosal epithelia have many important functions in protecting the host from pathogen invasion, as well as in establishing a symbiotic relationship with the human microbiota. Accordingly, mucosal epithelial cells and skin keratinocytes have specialized antimicrobial functions: for example, producing antimicrobial peptides, which limit the viability and multiplication of pathogens and symbiotic microorganisms that colonize these sites. The production of these antimicrobial molecules is induced by engagement of TLRs and NOD proteins and, presumably, other PRRs. Epithelial cells at the mucosal surface also produce mucins, which help to prevent the attachment and entry of pathogens.

Phagocytes

The phagocytic uptake of pathogens is crucial for host defence and is carried out by macrophages and neutrophils. These phagocytes are equipped with multiple antimicrobial mechanisms that are activated on initial contact with pathogens. They have a crucial role in defence against both intracellular bacteria and extracellular bacteria, as well as fungal pathogens. Phagocytosis is facilitated by opsonins, which are host products of the acute-phase response and the complement systems (discussed in the next section), through their ability to bind to both the cell walls of microorganisms and the opsonin receptors present on phagocytes.

Acute-phase proteins and complement

A variety of secreted proteins that function in the circulation and tissue fluids — acute-phase proteins and the complement system — constitute another module. Acute-phase proteins are secreted by hepatocytes in response to the pro-inflammatory cytokines IL-1β and IL-6, and the serum concentration of acute-phase proteins increases markedly at the early stages of infection. A key component of this response is the secreted PRRs: collectins, ficolins and pentraxins[41–43]. Their main functions are opsonizing microbial cells for phagocytosis and activating the complement system. Whereas collectins and ficolins initiate the lectin pathway of complement activation, pentraxins activate the classical pathway, which is also induced by antibodies[41–43]. Complement activation itself has several consequences, including the following: opsonization of pathogens, through the covalent attachment of C3 fragments; recruitment of phagocytes to the site of infection, through the release of proteolytic fragments of C4 and C5 that have chemotactic activity; and direct killing of pathogens, through the formation of the membrane-attack complex, which is the terminal component of the complement cascade[44].

Inflammasomes

Inflammasomes are protein complexes that activate pro-inflammatory caspases[25]. The activation of caspase 1, in particular, is required for processing the IL-1 family of cytokines, including IL-1β, IL-18 and IL-33. These complexes might also process proteins other than pro-inflammatory cytokines. Inflammasomes are activated by the NALP and NAIP subfamilies of NLRs (discussed earlier) in response to bacterial infections and some forms of cellular stress. IL-1-family cytokines have diverse functions in inflammation and host defence.

Natural killer cells

Natural killer (NK) cells are specialized in defence against intracellular pathogens, mainly viruses. These cells have two main functions: inducing the apoptosis of infected cells and producing cytokines, particularly IFN-γ. They express two types of receptor, activating and inhibitory, and these receptors recognize their cognate, host-encoded ligands on infected (target) cells[45]. The balance of expression of activating and inhibitory ligands by a target cell is thought to determine whether it is killed or spared by a particular NK cell. The mechanisms that control the production of these ligands are poorly understood but might involve cell-autonomous viral recognition by intracellular sensors of infection or cell-autonomous detection of excessive cellular stress. Recognition of viral infection by the infected cells themselves, through RIG-I or MDA5, and by plasmacytoid dendritic cells, through TLRs, also controls NK-cell activity, by eliciting the production of type I IFNs either directly or indirectly through the expression of IL-15. IL-15 also regulates NK-cell maintenance[46].

Type I IFNs and IFN-induced proteins

Type I IFNs and IFN-induced proteins have a crucial role in defence against viruses. Type I IFNs are produced in response to viral infections, and these proteins trigger the expression of more than 100 genes, the products of which have diverse antiviral activities[47]. Type-I-IFN production can be elicited in two ways: first, by intracellular sensors of infection (as described in the previous section); and, second, by TLR3, TLR7 and TLR9 (which are located intracellularly, on endosomes). The first mode of production is ubiquitous and occurs in virally infected cells. It results in autocrine or paracrine IFN-mediated signalling, which confers an antiviral state on the infected cell and neighbouring cells. By contrast, the second mode of production involves the engagement of TLR7 or TLR9, which results in specialized type-I-IFN-producing cells, known as plasmacytoid dendritic cells, producing systemic levels of IFN-α[48]. In almost all cases, type I IFNs are produced in response to viral or bacterial nucleic acids. The only exception to this seems to be IFN-β production in response to TLR4 ligands, which are not nucleic acids.

Eosinophils, basophils and mast cells

Eosinophils, basophils and their products form a host-defence module involved in protection against multicellular parasites, such as helminths. Mast cells are also a component of this module, although their function is not restricted to protection against parasites[49]. Mast cells reside in mucosal and connective tissues, whereas eosinophils and basophils are recruited to the sites of infection from the circulation. During bacterial infection, mast cells can be activated directly by TLRs[49]. The way in which parasites activate mast cells, basophils and eosinophils is largely unknown. Recently, however, one of the main parasite-associated cell-wall components, chitin, was found to induce eosinophil recruitment[50]. Interestingly, the main defensive strategy that components of this module use against parasites does not seem to target these pathogens directly (although direct effects do occur). Instead, it is the host tissues, particularly the mucosal epithelia, smooth muscles and vasculature, that are the main targets of the immune response. These tissues are affected by the mediators released by mast cells and basophils in a way that limits

the spread of parasites and promotes their expulsion from the host. The function of this module is regulated by several cytokines, including IL-4, IL-5, IL-9 and IL-13 (ref. 51).

Evolution and functional organization of the innate immune system

The various modules of the innate immune system evolved at different stages of phylogeny in response to specific challenges imposed by different classes of pathogen. The appearance of distinct innate host-defence modules during evolution also reflects changes in anatomy and physiology as animals evolved. In addition, a given module is not the same in different animals and might have been expanded or contracted during evolution in response to specific needs. The composition of the innate host-defence modules in any given animal species is therefore one of many possible configurations, presumably the optimal one for affording maximum protection in the specific physiological context. Thus, the innate immune system, although ancient in origin, is not equivalent in different animal phyla or classes or even between different species of the same class. For example, the number, expression and regulation of the antimicrobial molecules known as defensins varies between mammalian species.

More marked changes in the structure of the innate immune system can be seen at different stages of phylogeny. NK cells, type I IFNs, eosinophils and basophils are all unique to vertebrates. Moreover, the elimination of virally infected cells by NK cells is a viable defensive strategy only in complex metazoans, which have renewable tissues, and not in invertebrates, which consist of post-mitotic cells, most of which do not self-renew. (The extreme tissue and organ autonomy and regenerative capacity in plants might similarly explain why immune-response-associated cell death is one of the main host-defence strategies in plant immunity.) In addition, mast-cell-mediated and basophil-mediated defence against multicellular parasites is based, in part, on the effects of these cells on the vasculature, which is absent in invertebrates with an open circulatory system. Also, most arthropods are not suitable hosts for helminths because they are not large enough to accommodate them. Accordingly, the components of innate immunity against parasites are absent from most or all invertebrate phyla.

Another important aspect of the organization of the innate immune system is that modules that are co-induced by an infection tend to develop functional links and are usually co-regulated by the same inducible signals, most commonly cytokines. For example, phagocytes, the complement system and acute-phase proteins are functionally linked through opsonization-dependent phagocytosis, and these two modules are co-induced during many bacterial infections. A more specific example is that TLR-activated macrophages produce IL-6, which induces hepatocytes to secrete opsonins during the acute-phase response. Similarly, NK cells and type I IFNs are co-induced by viral infections and are functionally coupled. Not all modules of the innate immune system are co-induced by a given infection, however. The modules that are not functionally coupled (for example, NK cells and basophils) are triggered by distinct pathways and are not co-regulated by the same cytokines.

The modules of innate host defence are activated by primary sensors of infection, in most cases by PRRs (as discussed earlier). TLRs can activate multiple modules (mucosal epithelium, phagocytes, acute-phase proteins and type I IFNs), whereas other PRRs seem to be more specialized.

Innate control of adaptive immune responses

In addition to direct activation of innate host-defence mechanisms, some PRRs are coupled to the induction of adaptive immune responses. As discussed earlier, conventional lymphocytes (most αβ T cells and B2 cells) express antigen receptors with random specificities and therefore recognize antigens that lack any intrinsic characteristics indicative of their origin. Therefore, conventional lymphocytes require instructions indicating the origin of the antigen they recognize. These instructions come from the innate immune system in the form of specialized signals inducible by PRRs[4], which can sense infection because of their specificity for products of microbial origin. Therefore, the basic principle of innate control of adaptive immunity is based on establishing an association between the antigens recognized by lymphocytes and the microbial products (that is, PAMPs) recognized by PRRs.

For T cells, this association is interpreted by dendritic cells. Dendritic cells reside in most peripheral tissues, where they monitor the tissue

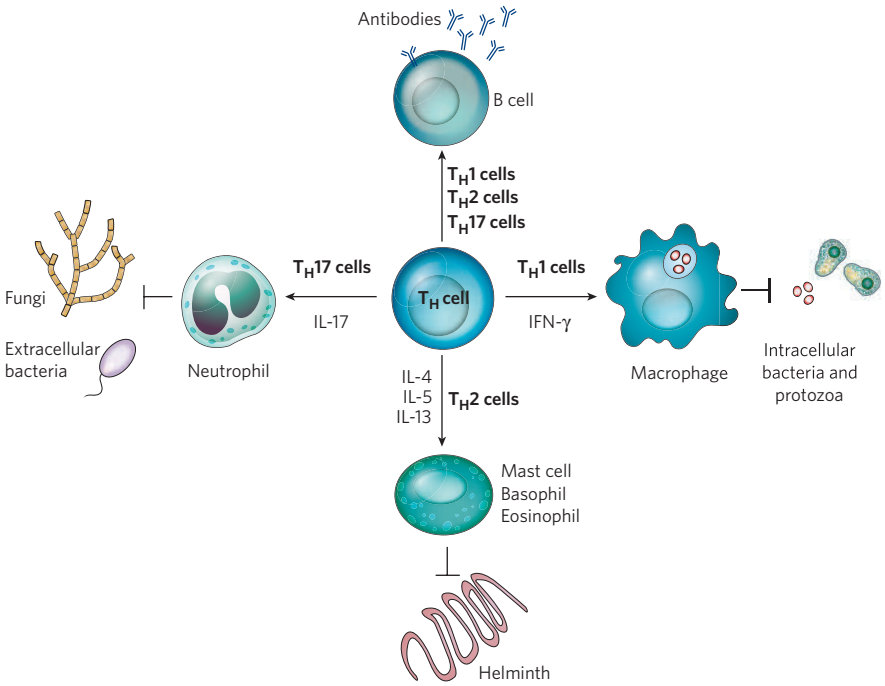


Figure 1 | Effector T_H-cell lineage and pathogen class. When circulating 'naive' T_H cells first recognize their cognate antigen, they differentiate into one of several effector-cell lineages (listed in bold), depending on the infecting pathogen. T_H1, T_H2 and T_H17 cells are the known types of effector T_H cell; however, other types of effector T_H cell probably exist. Each T_H-cell lineage is characterized by the cytokines that are produced and by the innate immune effector mechanism that is activated (denoted by arrows). It is possible (but has not been proved) that every module of the innate immune system is controlled by a dedicated effector T_H-cell lineage.

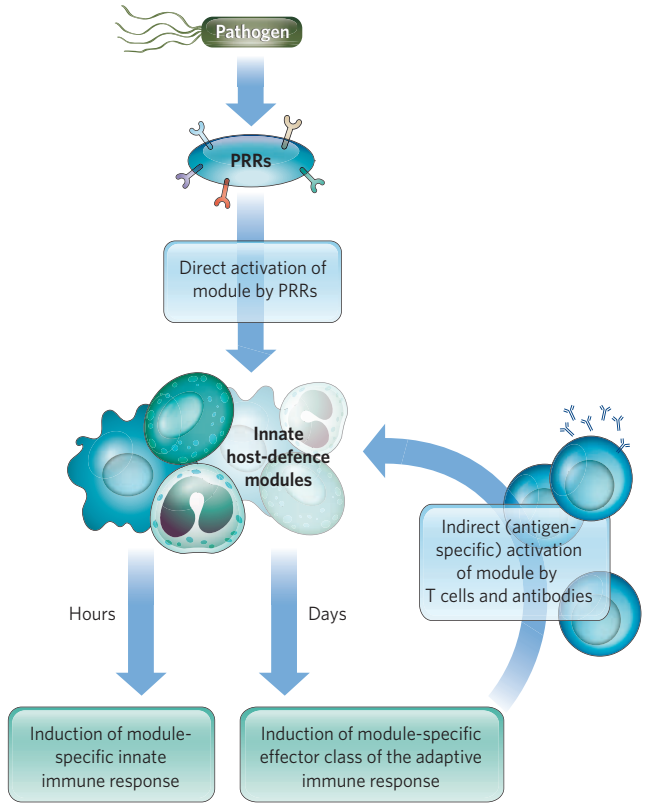


Figure 2 | Activation of host-defence mechanisms. Host-defence mechanisms can be induced directly, by engagement of PRRs, or indirectly, by T cells and/or antibodies. Each module is characterized by distinct antimicrobial defence mechanisms and can instruct the adaptive immune system to mount a response involving a module-specific effector class. After an adaptive immune response has been initiated, it results in antigen-specific activation of the same innate immune module that instructed the adaptive immune response. For example, macrophages can be activated either directly by TLRs or indirectly by T_H1 cells, through IFN-γ, CD40 ligand and other signals. Eosinophils can be activated either directly by an unidentified PRR or indirectly by T_H2 cells. And the classical pathway of complement activation can be induced either directly by pentraxins or indirectly by antibodies. Antigen-specific activation of the innate host-defence modules is more efficient than direct activation and is often required for pathogen clearance.

environment for the presence of pathogens by using various PRRs. When a pathogen is encountered by a dendritic cell, it is taken up by phagocytosis, and its protein constituents are processed into antigenic peptides, which are presented at the cell surface by MHC class I and/or class II molecules. For MHC class II molecules, the antigenic peptides selected for presentation derive from the phagosome in which the pathogen was internalized in response to the triggering of TLRs or other PRRs[52]. A similar mechanism might also operate for MHC class I molecules. Therefore, the association between an antigen and a PAMP is established as a result of their presence in the same phagocytosed 'cargo' (for example, a bacterial cell). PRRs also activate dendritic cells, inducing them to produce cytokines and express cell-surface signals and to migrate to the lymph nodes through the lymphatic vessels that drain the site of infection. When these dendritic cells reach the lymph nodes, they present the pathogen-derived antigens, together with PRR-induced signals (cytokines and cell-surface-associated molecules), to T cells. This results in T-cell activation and, in the case of T_H ($CD4^+$) cells, differentiation into one of several types of effector T_H cell[53].

For B cells, the association between an antigen and a PAMP can be established directly, when the two are physically linked in a single molecule or particle. This presumably occurs through co-engagement of a B-cell receptor and a PRR. In the extreme case, a TLR ligand (for example, lipopolysaccharide or flagellin) is itself recognized by the B-cell receptor and by a corresponding TLR expressed by a B cell. Antigens of this class, which combine ligands for both innate and adaptive immune recognition, are called T-independent antigens, because they can elicit B-cell responses without 'help' from T_H cells. When an antigen and a PAMP are not physically linked, their association is established through effector T_H cells that have previously been activated by dendritic cells. Antigens of this class (usually proteins) are called T-dependent antigens.

The antigen receptors of innate-like lymphocytes are skewed towards the recognition of microbial products, so the activation of these cells does not require the same elaborate mechanisms as for conventional lymphocytes. Indeed, B1 cells can be activated directly by PRRs and are programmed to produce antibodies with a broad specificity for common bacterial antigens[38]. Innate-like T cells recognize microbial antigens (such as lipids, glycolipids and formylpeptides) presented by non-classical MHC molecules. In certain cases, these cells recognize MHC-like molecules that do not seem to present any antigens but whose expression is inducible by PRRs. In such cases, the production of T-cell-receptor ligands in response to microbial products might be sufficient to signal the presence of infection.

T_H cells can differentiate into several types of effector cell: T_H1, T_H2 and T_H17 cells[54] (Fig. 1). These cells are characterized by the production of distinct sets of cytokines[55,56]. T_H1 cells produce IFN-γ and activate macrophages and other cell types to trigger defence against intracellular pathogens. T_H1-cell-derived IFN-γ also instructs B cells to produce antibodies of the IgG2 subclass. T_H2 cells are involved in protection against multicellular parasites and produce IL-4, IL-5 and IL-13 (ref. 51). These cytokines control the function of eosinophils, basophils and the mucosal epithelia. IL-4 also instructs B cells to produce antibodies of the IgE class, which are important in defence against parasites through their effects on mast-cell and basophil activation[57]. Finally, T_H17 cells produce IL-17, which induces non-haematopoietic cell types, including epithelial cells, to produce chemokines that recruit neutrophils to the site of infection[58]. T_H17-cell responses are involved in protection against extracellular bacteria and fungi[14]. The differentiation of naive T_H cells (which have not previously encountered their cognate antigen) into the three effector-cell lineages, T_H1, T_H2 and T_H17 cells, is controlled by transcriptional master regulators, in this case T-bet, GATA-binding protein 3 (GATA3) and retinoic-acid-receptor-related orphan receptor-γt (RORγt), respectively[59]. The expression of these master regulators is controlled by cytokines produced by antigen-presenting cells (such as dendritic cells) in response to PRR activation.

The effector response in each case is thus dictated by the innate immune system. In terms of T_H cells, TLR engagement induces IL-12 production, which directs T_H cells to differentiate into T_H1 cells. By contrast, TLR-induced IL-6, together with transforming growth factor-β (from an unknown cellular source), induces differentiation into T_H17 cells[58,59]. And dectin-1 engagement results in the production of IL-23, which is required for T_H17-cell function and/or maintenance[17,60]. The mechanisms of T_H2-cell generation are unknown but presumably follow a similar principle, with the dedicated cytokines likely to be IL-4 and thymic stromal lymphopoietin (TSLP), produced in response to engagement of an unidentified sensor after helminth infection[61]. For other cell types, type I IFNs (which are produced in response to TLR engagement or RIG-I, MDA5 or DAI engagement during viral infections) regulate the function of cytotoxic T cells and NK cells, either directly or indirectly by inducing IL-15 production[62].

Importantly, the adaptive immune response ultimately results in an antigen-specific activation of the effector mechanisms of the innate immune system. Thus, the effector T_H cells produce the appropriate effector cytokines that activate a specific module of the innate immune system (Fig. 2), including activation of macrophages by T_H1 cells, activation of neutrophils by T_H17 cells and activation of eosinophils, mast cells and basophils by T_H2 cells[54,63]. Similarly to NK cells, cytotoxic T cells induce apoptosis of infected cells, except that the T-cell response is antigen specific. Likewise, antibodies activate the modules of the innate immune system in a class-dependent (and antigen-dependent)

manner. IgG activates complement and opsonizes pathogens to aid their phagocytosis by macrophages and neutrophils, whereas IgE activates mast cells and basophils. Each of the innate effector responses can therefore be activated either directly, by the appropriate PRRs at the early stages of infection, or indirectly, by T cells and antibodies (in an antigen-specific manner) at the later, effector, stages of the immune response (Fig. 2). Furthermore, each effector mechanism of the adaptive immune system might have evolved to activate the appropriate host-defence module of the innate immune system.

The relative contributions of the innate immune system and the adaptive immune system during bacterial infections have been investigated extensively. One important principle that has emerged from these studies is that, although innate host defence is crucial for controlling an infection, it is often insufficient for pathogen clearance[64]. For example, to clear a *Listeria monocytogenes* infection requires functional T-cell responses[64]. It therefore seems that the innate immune system in vertebrates evolved to depend, to some extent, on antigen-specific (adaptive) immunity. This might explain why the mammalian innate immune system, unlike that of arthropods, is not self-sufficient at affording protection against many infections. It should be noted, however, that our understanding of host defence might be biased because almost all studies are based on symptomatic infections. Asymptomatic infections are presumably common, and many of these infections might be cleared efficiently by innate host-defence mechanisms.

Conclusions and perspectives

The adaptation of microorganisms to host niches can benefit the host, as is the case with many symbiotic microorganisms. In some cases, however, adaptation negatively affects tissue physiology and might directly damage tissues, resulting in symptomatic infectious disease. The symptoms of an infectious disease can also be caused by excessive immune and inflammatory responses, and these can often be more damaging to the host than the virulence activity of the pathogen that elicited them. Thus, it is just as crucial for the host to limit the immunopathology as it is to protect against the infecting pathogen. The trade-off between immunopathology and protection against infection has presumably resulted in an optimal balance of the sensitivity and intensity of the immune response. This balance is probably not hard-wired, because the immune response needs to vary in intensity and duration depending on the infecting pathogen. One possible solution to the conundrum of how the balance is achieved would be that the tissue damage caused directly by the pathogen can be distinguished from the damage inflicted by the immune response. If this is the case and the two types of tissue damage are differentially detected by the host, then the extent of the damage might negatively control the intensity and duration of the immune response. Understanding the balance between the two conflicting causes of infectious-disease symptoms is crucial for the development of appropriate therapeutic strategies[65].

Another area of great importance is understanding the principles of protective immunity. What matters to the host organism is not the induction of an immune response but whether the immune response protects against a given infection. Not all immune responses are protective. Only immune responses of the correct effector class directed at particular antigens can provide protection against infection. It is generally assumed that the effector response depends on the pathogen type. However, in most cases, the crucial pathogen features that determine the effector immune response are unclear. Similarly, it is not known how the relevant features of pathogens are translated into the appropriate set of signals that determine the effector response. Understanding these principles is vital for the development of vaccines that can elicit protective immunity. ■

1. Hacker, J. & Carniel, E. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* **2,** 376–381 (2001).
2. Gal-Mor, O. & Finlay, B. B. Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell. Microbiol.* **8,** 1707–1719 (2006).
3. Finlay, B. B. & Falkow, S. Common themes in microbial pathogenicity revisited. *Microbiol. Mol. Biol. Rev.* **61,** 136–169 (1997).
4. Janeway, C. A. Jr. Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb. Symp. Quant. Biol.* **54,** 1–13 (1989).
5. Schatz, D. G., Oettinger, M. A. & Schlissel, M. S. V(D)J recombination: molecular biology and regulation. *Annu. Rev. Immunol.* **10,** 359–383 (1992).
6. Rifkin, I. R., Leadbetter, E. A., Busconi, L., Viglianti, G. & Marshak-Rothstein, A. Toll-like receptors, endogenous ligands, and systemic autoimmune disease. *Immunol. Rev.* **204,** 27–42 (2005).
7. Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. & Medzhitov, R. Recognition of commensal microflora by Toll-like receptors is required for intestinal homeostasis. *Cell* **118,** 229–241 (2004).
8. Akira, S., Uematsu, S. & Takeuchi, O. Pathogen recognition and innate immunity. *Cell* **124,** 783–801 (2006).
9. Thoma-Uszynski, S. *et al.* Induction of direct antimicrobial activity through mammalian Toll-like receptors. *Science* **291,** 1544–1547 (2001).
10. Liu, P. T., Krutzik, S. R. & Modlin, R. L. Therapeutic implications of the TLR and VDR partnership. *Trends Mol. Med.* **13,** 117–124 (2007).
11. Liu, P. T. *et al.* Toll-like receptor triggering of a vitamin D-mediated human antimicrobial response. *Science* **311,** 1770–1773 (2006).
12. Brown, G. D. & Gordon, S. Immune recognition. A new receptor for β-glucans. *Nature* **413,** 36–37 (2001).
13. Robinson, M. J., Sancho, D., Slack, E. C., LeibundGut-Landmann, S. & Reis e Sousa, C. Myeloid C-type lectins in innate immunity. *Nature Immunol.* **7,** 1258–1265 (2006).
14. LeibundGut-Landmann, S. *et al.* Syk- and CARD9-dependent coupling of innate immunity to the induction of T helper cells that produce interleukin 17. *Nature Immunol.* **8,** 630–638 (2007).
15. Saijo, S. *et al.* Dectin-1 is required for host defense against *Pneumocystis carinii* but not against *Candida albicans*. *Nature Immunol.* **8,** 39–46 (2007).
16. Taylor, P. R. *et al.* Dectin-1 is required for β-glucan recognition and control of fungal infection. *Nature Immunol.* **8,** 31–38 (2007).
17. Brown, G. D. Dectin-1: a signalling non-TLR pattern-recognition receptor. *Nature Rev. Immunol.* **6,** 33–43 (2006).
18. Fritz, J. H., Ferrero, R. L., Philpott, D. J. & Girardin, S. E. Nod-like proteins in immunity, inflammation and disease. *Nature Immunol.* **7,** 1250–1257 (2006).
19. Inohara, N., Chamaillard, M., McDonald, C. & Nunez, G. NOD-LRR proteins: role in host–microbial interactions and inflammatory disease. *Annu. Rev. Biochem.* **74,** 355–383 (2005).
20. Meylan, E., Tschopp, J. & Karin, M. Intracellular pattern recognition receptors in the host response. *Nature* **442,** 39–44 (2006).
21. Ting, J. P. & Davis, B. K. CATERPILLER: a novel gene family important in immunity, cell death, and diseases. *Annu. Rev. Immunol.* **23,** 387–414 (2005).
22. Fritz, J. H. *et al.* Nod1-mediated innate immune recognition of peptidoglycan contributes to the onset of adaptive immunity. *Immunity* **26,** 445–459 (2007).
23. Kobayashi, K. S. *et al.* Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract. *Science* **307,** 731–734 (2005).
24. Eckmann, L. & Karin, M. NOD2 and Crohn's disease: loss or gain of function? *Immunity* **22,** 661–667 (2005).
25. Martinon, F. & Tschopp, J. Inflammatory caspases: linking an intracellular innate immune system to autoinflammatory diseases. *Cell* **117,** 561–574 (2004).
26. Yoneyama, M. *et al.* The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nature Immunol.* **5,** 730–737 (2004).
27. Takaoka, A. *et al.* DAI (DLM-1/ZBP1) is a cytosolic DNA sensor and an activator of innate immune response. *Nature* **448,** 501–505 (2007).
28. Gitlin, L. *et al.* Essential role of mda-5 in type I IFN responses to polyriboinosinic: polyribocytidylic acid and encephalomyocarditis picornavirus. *Proc. Natl Acad. Sci. USA* **103,** 8459–8464 (2006).
29. Hornung, V. *et al.* 5′-Triphosphate RNA is the ligand for RIG-I. *Science* **314,** 994–997 (2006).
30. Kato, H. *et al.* Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* **441,** 101–105 (2006).
31. Pichlmair, A. *et al.* RIG-I-mediated antiviral responses to single-stranded RNA bearing 5′-phosphates. *Science* **314,** 997–1001 (2006).
32. Seth, R. B., Sun, L. & Chen, Z. J. Antiviral innate immunity pathways. *Cell Res.* **16,** 141–147 (2006).
33. Ishii, K. J. *et al.* A Toll-like receptor-independent antiviral response induced by double-stranded B-form DNA. *Nature Immunol.* **7,** 40–48 (2006).
34. Okabe, Y., Kawane, K., Akira, S., Taniguchi, T. & Nagata, S. Toll-like receptor-independent gene induction program activated by mammalian DNA escaped from apoptotic DNA degradation. *J. Exp. Med.* **202,** 1333–1339 (2005).
35. Stetson, D. B. & Medzhitov, R. Recognition of cytosolic DNA activates an IRF3-dependent innate immune response. *Immunity* **24,** 93–103 (2006).
36. Honda, K., Takaoka, A. & Taniguchi, T. Type I interferon gene induction by the interferon regulatory factor family of transcription factors. *Immunity* **25,** 349–360 (2006).
37. Bendelac, A., Bonneville, M. & Kearney, J. F. Autoreactivity by design: innate B and T lymphocytes. *Nature Rev. Immunol.* **1,** 177–186 (2001).
38. Berland, R. & Wortis, H. H. Origins and functions of B-1 cells with notes on the role of CD5. *Annu. Rev. Immunol.* **20,** 253–300 (2002).
39. Hansen, T. H., Huang, S., Arnold, P. L. & Fremont, D. H. Patterns of nonclassical MHC antigen presentation. *Nature Immunol.* **8,** 563–568 (2007).
40. Janeway, C. A. Jr, Jones, B. & Hayday, A. Specificity and function of T cells bearing γδ receptors. *Immunol. Today* **9,** 73–76 (1988).
41. Bottazzi, B. *et al.* Pentraxins as a key component of innate immunity. *Curr. Opin. Immunol.* **18,** 10–15 (2006).
42. Endo, Y., Takahashi, M. & Fujita, T. Lectin complement system and pattern recognition. *Immunobiology* **211,** 283–293 (2006).
43. Holmskov, U., Thiel, S. & Jensenius, J. C. Collectins and ficolins: humoral lectins of the innate immune defense. *Annu. Rev. Immunol.* **21,** 547–578 (2003).
44. Carroll, M. C. & Fischer, M. B. Complement and the immune response. *Curr. Opin. Immunol.* **9,** 64–69 (1997).
45. Lanier, L. L. NK cell recognition. *Annu. Rev. Immunol.* **23,** 225–274 (2005).

46. Ma, A., Koka, R. & Burkett, P. Diverse functions of IL-2, IL-15, and IL-7 in lymphoid homeostasis. *Annu. Rev. Immunol.* **24,** 657–679 (2006).
47. Hiscott, J. Triggering the innate antiviral response through IRF-3 activation. *J. Biol. Chem.* **282,** 15325–15329 (2007).
48. Liu, Y. J. IPC: professional type 1 interferon-producing cells and plasmacytoid dendritic cell precursors. *Annu. Rev. Immunol.* **23,** 275–306 (2005).
49. Galli, S. J., Nakae, S. & Tsai, M. Mast cells in the development of adaptive immune responses. *Nature Immunol.* **6,** 135–142 (2005).
50. Reese, T. A. *et al.* Chitin induces accumulation in tissue of innate immune cells associated with allergy. *Nature* **447,** 92–96 (2007).
51. Stetson, D. B. *et al.* T_H2 cells: orchestrating barrier immunity. *Adv. Immunol.* **83,** 163–189 (2004).
52. Blander, J. M. & Medzhitov, R. Toll-dependent selection of microbial antigens for presentation by dendritic cells. *Nature* **440,** 808–812 (2006).
53. Banchereau, J. & Steinman, R. M. Dendritic cells and the control of immunity. *Nature* **392,** 245–252 (1998).
54. Reinhardt, R. L., Kang, S. J., Liang, H. E. & Locksley, R. M. T helper cell effector fates — who, how and where? *Curr. Opin. Immunol.* **18,** 271–277 (2006).
55. Glimcher, L. H. & Murphy, K. M. Lineage commitment in the immune system: the T helper lymphocyte grows up. *Genes Dev.* **14,** 1693–1711 (2000).
56. Seder, R. A. & Paul, W. E. Acquisition of lymphokine-producing phenotype by $CD4^+$ T cells. *Annu. Rev. Immunol.* **12,** 635–673 (1994).
57. Nelms, K., Keegan, A. D., Zamorano, J., Ryan, J. J. & Paul, W. E. The IL-4 receptor: signaling mechanisms and biologic functions. *Annu. Rev. Immunol.* **17,** 701–738 (1999).
58. Weaver, C. T., Hatton, R. D., Mangan, P. R. & Harrington, L. E. IL-17 family cytokines and the expanding diversity of effector T cell lineages. *Annu. Rev. Immunol.* **25,** 821–852 (2007).
59. Reiner, S. L. Development in motion: helper T cells at work. *Cell* **129,** 33–36 (2007).
60. Acosta-Rodriguez, E. V. *et al.* Surface phenotype and antigenic specificity of human interleukin 17-producing T helper memory cells. *Nature Immunol.* **8,** 639–646 (2007).
61. Liu, Y. J. *et al.* TSLP: an epithelial cell cytokine that regulates T cell differentiation by conditioning dendritic cell maturation. *Annu. Rev. Immunol.* **25,** 193–219 (2007).
62. Nguyen, K. B. *et al.* Coordinated and distinct roles for IFN-αβ, IL-12, and IL-15 regulation of NK cell responses to viral infection. *J. Immunol.* **169,** 4279–4287 (2002).
63. Shinkai, K., Mohrs, M. & Locksley, R. M. Helper T cells regulate type-2 innate immunity *in vivo*. *Nature* **420,** 825-829 (2002).
64. Unanue, E. R. Studies in listeriosis show the strong symbiosis between the innate cellular system and the T-cell response. *Immunol. Rev.* **158,** 11–25 (1997).
65. Ewald, P. W. *Evolution of Infectious Disease* (Oxford Univ. Press, Oxford, 1996).

Acknowledgements I thank I. Brodsky and A. Iwasaki for critical reading of the manuscript. I apologize to the many authors whose work could not be cited directly because of space limitations.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to the author (ruslan.medzhitov@yale.edu).

Manipulation of host-cell pathways by bacterial pathogens

Amit P. Bhavsar[1], Julian A. Guttman[1,2] & B. Brett Finlay[1]

Bacterial pathogens operate by attacking crucial intracellular pathways in their hosts. These pathogens usually target more than one intracellular pathway and often interact at several points in each of these pathways to commandeer them fully. Although different bacterial pathogens tend to exploit similar pathway components in the host, the way in which they 'hijack' host cells usually differs. Knowledge of how pathogens target distinct cytoskeletal components and immune-cell signalling pathways is rapidly advancing, together with the understanding of bacterial virulence at a molecular level. Studying how these bacterial pathogens subvert host-cell pathways is central to understanding infectious disease.

Infectious diseases are major threats to human health worldwide, and tremendous effort has gone into understanding various infectious agents and their mechanisms of virulence. One theme that emerged early from studies of bacterial pathogens is that many inject pathogenic factors (also known as effectors or virulence factors) directly into host cells as part of their pathogenic strategy, and these factors specifically target crucial intracellular pathways in the host. This pathogenic strategy was shown to be used by certain extracellular and intracellular pathogens. These findings led to an exciting and dynamic crossover between related disciplines, including microbiology, cell biology, biochemistry and immunology. And from such interdisciplinary studies arose the central tenet that bacterial pathogens specifically attack key intracellular-signalling and cytoskeletal pathways to alter host responses in a way that favours the pathogen. Another important finding of these studies was that the same pathways are often targeted by different bacterial effectors and that these pathways can be attacked at several points by a single pathogen, ensuring an override of important cellular functions.

In this review, we provide an overview of the current understanding of how bacterial pathogens interact with host cells. How these microorganisms exploit host cells is discussed in terms of both promoting the bacterial life cycle and evading the host immune response. We focus on examples of pathogenic bacteria that interact with mammalian intracellular-signalling, vesicular-trafficking and cytoskeletal pathways, because these are some of the best studied or most rapidly advancing areas.

Cell biology of bacterial infection

Bacterial pathogens use a range of effectors to subvert and control normal cellular functions. Effectors are usually specialized proteins that are injected directly into the cytosol of the host cell by a type III secretion system (T3SS) or a type IV secretion system (T4SS). These secretion systems consist of a structurally conserved proteinaceous apparatus that is shaped like a needle[1].

The concept of secreted proteins functioning as agents of microbial virulence is not new. Toxins have been recognized in this capacity for decades. But the ability to inject effectors directly into mammalian or plant cells is repeatedly encountered when considering the cell biology of the infectious process. Therefore, it is an important process in microbial pathogenesis as it is understood at present, and an important interface between pathogens and their hosts.

To promote their life cycle, bacterial pathogens can use host cells to aid their own adherence, replication and/or dissemination. The initial step in colonization by bacteria is adherence. Bacterial pathogens have a large variety of cell-surface adhesins, including fimbriae and afimbrial adhesins (see ref. 2 for a review), that enable them to attach to host cells. Some of these adhesins also have a further role: they bind to their cognate receptors on non-phagocytic cells, thereby allowing bacteria to be taken up by these cells. Such adhesins with dual roles include the invasins of *Yersinia* spp. and the internalins of *Listeria* spp.[3]. The internalization mechanisms of intracellular pathogens differ on the basis of the effector involved, and they require that complementary host intracellular-signalling pathways are commandeered.

The most commonly described cellular target of pathogens is the cytoskeleton. Various intracellular microorganisms harness cytoskeletal components to gain entry to, and to propel themselves within, host cells (see ref. 4 for a review) (Fig. 1). The cytoskeleton of eukaryotic cells is composed of actin filaments, microtubules and intermediate filaments. In terms of bacterial pathogenesis, the most extensively studied of these are actin filaments. Bacterial pathogens do not usually interact directly with actin filaments themselves. Instead, they subvert and control the polymerization of actin filaments by modulating cellular regulators of this process, such as small Rho-like G proteins[5], through the action of delivered effectors (Fig. 1a).

Some bacterial pathogens remain in vacuoles after internalization, and these microorganisms often use effectors to modulate vesicular trafficking, providing a protective niche within host cells (Fig. 1c) (including in macrophages and neutrophils, which normally kill bacteria)[6]. In addition, pathogens can interact with cell-death pathways (including apoptosis), modulating host-cell death to facilitate pathogen survival in the host[7]. Moreover, one of the key ways that pathogens evade or subvert the host immune response (both the innate and the adaptive immune mechanisms)[8] is to secrete effector proteins (Fig. 1d). These steps of the colonization and immune-evasion processes are discussed in more detail later.

Manipulation of the cytoskeleton and membranous structures

How pathogenic bacteria exploit the host-cell cytoskeleton, membranous structures and key signalling pathways to their advantage is discussed in this section, together with the strategies and rationales these

[1]The University of British Columbia, Michael Smith Laboratories, 301-2185 East Mall, Vancouver, British Columbia V6T 1Z4, Canada. [2]Present address: Simon Fraser University, 8888 University Drive, Department of Biological Sciences, Room B8276, Shrum Science Centre, Burnaby, British Columbia V5A 1S6, Canada.

microorganisms use to invade, survive intracellularly and replicate in the host.

Interactions of bacteria with the actin cytoskeleton

Bacterial pathogens manipulate the cytoskeleton to help invade a host cell and/or to gain motility in the cell, as mentioned earlier. They often interact with actin filaments in particular, and they do so by modulating G proteins. This process is exemplified by the interaction of the invasive bacterium *Salmonella enterica* with mammalian cells. During this process, *S. enterica* delivers the T3SS effector proteins SopE and SopE2 into the host cell. These effectors function as guanine-nucleotide-exchange factors for G proteins, activating the G protein CDC42 and the RAC family of G proteins in the target cell[9–11]. This G-protein activation, in turn, induces the generation of actin-rich membrane ruffles that engulf and internalize the bacteria (Fig. 1a). An interesting alternative strategy has recently been reported[12]: bacterial effector proteins that contain a Trp-X-X-X-Glu motif suppress the signalling of active G proteins and mimic these active G proteins themselves, thereby obviating the need for modulating the GTPase activity of G proteins (Box 1).

After invasion and escape from membrane-enclosed vesicles into the cytosol, many pathogens also manipulate actin-filament dynamics so that they can move within the infected host cell (see ref. 4 for a review). They do so by recruiting actin to just one of their poles, through bacterial-protein-mediated nucleation of actin. For example, the intracellular motility of *Shigella flexneri* is mediated by the bacterial effector IcsA. IcsA interacts directly with the host protein N-WASP (neural Wiskott–Aldrich syndrome protein; also known as WASL), which in turn recruits a complex known as the Arp2/3 complex (consisting of seven host proteins, including actin-related protein 2 (ARP2) and ARP3).This complex polymerizes actin filaments behind the advancing bacterium[13]. By contrast, the cytosolic motility of *Listeria* spp. is mediated by the bacterial protein ActA, which binds directly to both the Arp2/3 complex and the actin-associated protein VASP (vasodilator-stimulated phosphoprotein)[14,15].

The hijacking of actin-associated cytoskeletal components can also occur during infection with extracellular pathogens. For example, the attaching and effacing human pathogens enterohaemorrhagic *Escherichia coli* (EHEC) and enteropathogenic *E. coli* (EPEC) have an elaborate actin-recruiting process (Fig. 2). In this case, the bacterial effector protein Tir mediates extensive modification of host-cell actin filaments beneath the adherent microorganism. Tir is delivered into the target cell by the T3SS and embeds itself in the plasma membrane, where it anchors the bacterium firmly by binding to the bacterial outer-membrane protein intimin. In the case of EPEC infection, Tir is tyrosine phosphorylated on the cytoplasmic face of the host-cell plasma membrane and recruits the host adaptor protein Nck (non-catalytic region of tyrosine kinase)[16]. During EHEC infection, Nck is not recruited; instead, an additional bacterial effector, $EspF_u$ (also known as TccP), is involved[17,18]. Downstream of this protein (Nck or $EspF_u$), N-WASP and the Arp2/3 complex are recruited, and they mediate the polymerization of actin filaments beneath the adherent extracellular bacterium[16–18]. This results in the formation of a 'pedestal' on the host-cell

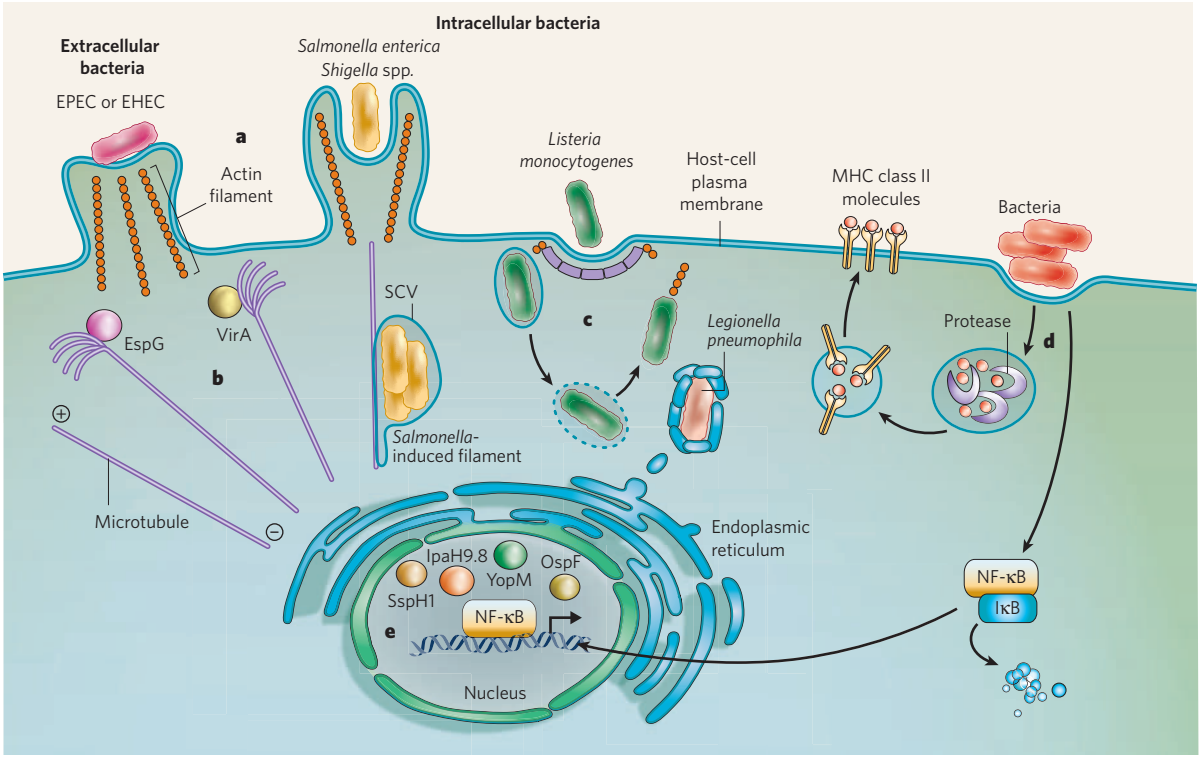


Figure 1 | The cell biology of bacterial infections. Both extracellular bacterial pathogens and intracellular bacterial pathogens initially interact with the plasma membrane of host cells. These pathogens commandeer various common structures and pathways. **a**, Extracellular pathogens (for example, EPEC and EHEC) routinely manipulate the actin cytoskeleton to generate actin-rich pedestal structures. By contrast, the invasive intracellular pathogens *Salmonella enterica* and *Shigella* spp. use this cytoskeletal component to invade the host cell. Another invasive intracellular pathogen, *Listeria monocytogenes*, uses clathrin-mediated endocytosis for invasion. **b**, During some bacterial infections, microtubules, another component of the cytoskeleton, are disassembled through the actions of particular effector proteins. Such proteins present during EPEC infection (EspG) and *Shigella flexneri* infection (VirA) disassemble microtubules in the vicinity of bacterial contact. **c**, After invasive bacteria have entered the host cell, they occupy a vacuole. The vacuole can offer protection against immune detection and can be a replicative niche (in the case of *S. enterica*). Alternatively, the bacteria can escape the vacuole and gain the ability to propel themselves through the cytosol (in the case of *L. monocytogenes* and *S. flexneri*). Organellar components can also be acquired: for example, *Legionella pneumophila*, which is usually internalized by phagocytosis, interacts with endoplasmic-reticulum-derived vesicles. **d**, Bacterial pathogens also have other effector-protein-based mechanisms for evading both the innate immune response and the adaptive immune response. Examples include subverting the presentation of bacterial antigen at the cell surface (adaptive) and interfering with the translocation of the pro-inflammatory transcriptional activator NF-κB to the cell nucleus (innate). **e**, The nucleus is also the destination of several T3SS effectors (for example, *S. enterica* SspH1, *Yersinia* spp. YopM and *S. flexneri* OspF and IpaH9.8), and the effects of this are beginning to come to light.

surface, where the pathogen resides. These pedestals can move ('surf') on the cell surface[19]: the actin-disassembly proteins cofilin and gelsolin have been identified in pedestals, and these proteins presumably regulate actin-filament dynamics in conjunction with the actin-assembly protein profilin. Numerous other actin-associated proteins — including cortactin[20], GRB2 (growth-factor-receptor-bound protein 2), LPP (lipoma-preferred partner), SHC (SRC-homology-2-domain-containing transforming protein C), vinculin and zyxin[21] — have been found in pedestals, although their precise organization during pedestal generation has not yet been determined. The protein α-actinin has also been identified in EPEC-induced pedestals, where it specifically interacts with the amino terminus of Tir[22]. Surprisingly, the endocytosis-associated protein dynamin has been found in EPEC-induced pedestals[23], as have the intermediate-filament proteins cytokeratin 8 and cytokeratin 18 (ref. 24) and the tight-junction component ZO1 (also known as TJP1)[25]. Therefore, pedestals are useful sites at which to study the interplay of cytoskeletal systems with components of signalling pathways, endocytosis pathways and intercellular junctions.

Interactions of microtubules with effectors

Microtubules are also commonly targeted by microorganisms. These polarized structures are normally used for structural support and as tracks to guide and transport intracellular cargo, with the aid of microtubule-associated molecular-motor proteins. During certain infections, both the cargo transport and the microtubule assembly and/or disassembly dynamics can be modified and controlled by the pathogen. For example, on invasion by *Shigella* spp., the VirA protein interacts directly with heterodimers of α-tubulin and β-tubulin, promoting destabilization of the microtubules[26] (Fig. 1b). This results in a localized absence of microtubules near the invading bacteria, thus aiding invasion by *Shigella* spp. A similar phenotype is seen during EPEC infections (Fig. 1b). In this case, localized microtubule depolymerization depends on the bacterial effector EspG, which, similarly to VirA, interacts directly with tubulin[27].

Whereas *Shigella* spp. and pathogenic *E. coli* disassemble microtubules, a strain of *Campylobacter jejuni* has been shown to use microtubules and their associated molecular motors to aid invasion. Microtubules are polar structures that have distinct fast-growing (plus) and slow-growing (minus) ends (Fig. 1b), allowing the directional transport of cargo in cells. There are two general types of microtubule-based molecular motor found in the host-cell cytosol: kinesins and cytoplasmic dynein. Members of the kinesin superfamily generally transport cargo towards the plus ends of microtubules. By contrast, cytoplasmic dynein is thought to be a minus-end-directed motor. Microtubules, and particularly dynein, have been implicated in invasion by a strain of *C. jejuni*[28]. Given that the minus ends of microtubules in cultured, non-polarized cells are directed towards the interior of the cell (Fig. 1b), this model seems plausible. However, in polarized cells (such as those present in intestinal epithelial barriers), microtubule polarity is reversed, so dynein would not transport *C. jejuni* towards the cell interior. Therefore, further investigation into the uptake of *C. jejuni* by host cells is required.

Life in a host cell

After entry to host cells, invasive pathogens are either localized in the cytosol or sequestered in vesicular structures. Presumably, all intracellular pathogens occupy a membrane-enclosed compartment at some point of their intracellular phase, even if only transiently. The initial compartments after internalization (vacuoles and modified phagosomes) are composed of membranous host-cell components; therefore, internalization often generates protected areas. Pathogens have adopted various strategies to multiply in, or escape from, these structures before surviving in the cytosol and then being disseminated throughout the host.

The ability to occupy a protected intracellular niche contributes to the pathogenesis of both *S. enterica* and *Legionella* spp. (see refs 29 and 30 for reviews). On passive internalization by phagocytic cells, *Legionella* spp. occupy a compartment known as a *Legionella*-containing vacuole (LCV). This phagosome is modified by the Dot/Icm secretion system (a T4SS).

Box 1 | Novel biochemical activities of translocated bacterial effectors

One of the central themes of bacterial pathogenesis is the manipulation of host-cell cytoskeletal components by injected effector proteins, which mediate this effect by subverting host G-protein signalling through their guanine-nucleotide-exchange and GTPase-activating activities. However, new data show that bacterial effector proteins can also catalyse novel, diverse and ingenious biochemical reactions that contribute to pathogenesis. For example, several effector proteins contain a Trp-X-X-X-Glu amino-acid motif, which enables them to 'mimic' the function of active (GTP-bound) G proteins, allowing downstream signalling and cytoskeletal remodelling[12]. This novel biochemical activity, the mechanism of which remains unknown, bypasses the requirement for G proteins.

Insights into the biochemistry of effector proteins that target host-cell ubiquitylation pathways are also emerging. The *Shigella flexneri* effector IpaH9.8 and the *Salmonella enterica* effector SspH1 were recently shown to have E3 ubiquitin–protein ligase activity[66]. IpaH9.8 catalyses the transfer of ubiquitin to the yeast mitogen-activated protein kinase (MAPK)-signalling-cascade member Ste7, presumably inducing its degradation and abrogating MAPK signalling. SspH1 catalyses the transfer of ubiquitin to the mammalian protein kinase PKN1, although the effect on PKN1-mediated signalling is unclear. By contrast, the *S. enterica* effector SseL was recently shown to catalyse the removal of polyubiquitin chains that had been attached to host proteins during infection[67]. It is intriguing to consider that the opposing biochemical activities of SspH1 and SseL might reflect the necessity of coordinating effector functions (discussed later).

MAPK-signalling pathways have an important role in immunity of the host to bacterial pathogens. Consequently, these canonical phosphorylation cascades are subject to attack by multiple bacterial effector proteins. Recently, two novel and diverse biochemical activities were identified for effectors targeting this pathway. Proteomic analyses indicate that the *Yersinia* spp. effector YopP/J is an acetylase[68]. The authors of this study propose that the transfer of acetyl moieties to key residues on MAPK substrates competes effectively with phosphorylation at these sites, thereby blocking signal transduction (Fig. 3). By contrast, the *S. flexneri* effector OspF has been shown to irreversibly dephosphorylate specific MAPKs — extracellular-signal-regulated kinase 2 (ERK2; also known as MAPK1), p38 MAPK (also known as MAPK14) and Jun amino-terminal kinase (JNK; also known as MAPK8) — by an elimination reaction that chemically modifies the key threonine residue of the substrate so that this MAPK cannot function in the signalling pathway[69] (Fig. 3). This enzymatic activity is called phosphothreonine-lyase activity, and it also seems to be present in other pathogenic bacteria. Interestingly, the authors of this study found that OspF had no phosphotyrosine-phosphatase activity, in contrast to another report that claimed OspF was a dual-specificity phosphatase[48]. The dephosphorylation of phosphotyrosine by an elimination reaction is a highly improbable mechanism, so to determine whether OspF is a dual-specificity phosphatase requires further study. Nevertheless, the identification of irreversible phosphothreonine-lyase activity opens the door to debate about the benefits for the infecting bacteria of irreversible modification of host-cell biology compared with those of reversible modulation.

T4SS effectors enable LCVs both to evade common phagocytic-degradation pathways (by preventing the acidification of vacuoles and the association of proteins found in late endosomes and lysosomes with LCVs) and to acquire components commonly found in secretory pathways. One of these acquired components is the GTPase ADP-ribosylation factor 1 (ARF1), the function of which is mediated by the bacterial effector RalF[31]. *Legionella* spp. further modify the phagosome by using the T4SS effector SidJ to recruit small endoplasmic-reticulum-derived vesicles to the phagosomal membrane, and then mediate fusion with these vesicles[32], potentially providing a nutrient-rich resource for the bacteria (Fig. 1c). In addition, the GTPase RAB1 is found at the LCV membrane and has a role in the fusion of endoplasmic-reticulum-derived vesicles with the LCV. The function of RAB1 is controlled by the bacterial effector DrrA (also known as SidM),

which has guanine-nucleotide-exchange-factor activity for RAB1 during infection with *Legionella* spp.[33]. As the infection progresses, these vesicles disappear (as assessed by morphological characteristics), and ribosomes are found to interact with the LCV membrane, thus placing the bacteria within a rough-endoplasmic-reticulum-like vacuole.

Similarly, *S. enterica* modifies its phagosome-like vacuole, by using a set of T3SS effectors, to provide a protective niche where the bacteria survive and replicate[30] (Fig. 1c). It accomplishes this by selectively interacting with components of the endocytic machinery of the host cell, thereby acquiring molecules such as early endosome antigen 1 and lysosomal-associated membrane protein 1 (refs 34, 35). Although it has long been accepted that lysosomes are inhibited from fusion with *Salmonella*-containing vacuoles (SCVs), recent advances have shown that lysosomes can readily fuse with SCVs during *S. enterica* infection[36], raising numerous questions about the exact mechanisms used to evade destruction by the host.

Active invasion by *S. flexneri* produces a vacuole around the invading microorganism. However, during invasion of epithelial cells, *S. flexneri* occupies this vacuole only briefly. This escape from the vacuole allows the bacterium to replicate in the host-cell cytosol and, eventually, to spread from cell to cell[37]. *Listeria monocytogenes*, a pathogen that is internalized through clathrin-mediated endocytosis[38], is also initially found in a vacuole (Fig. 1c). In a similar manner to invasion by *S. flexneri*, these vacuoles are short lived. *L. monocytogenes* uses the membrane-pore-forming toxin listeriolysin O, as well as the enzymes PlcA and PlcB, to destroy the surrounding membrane, thereby allowing escape from the vacuole and, subsequently, replication within the host-cell cytosol and actin-mediated spreading from cell to cell[39].

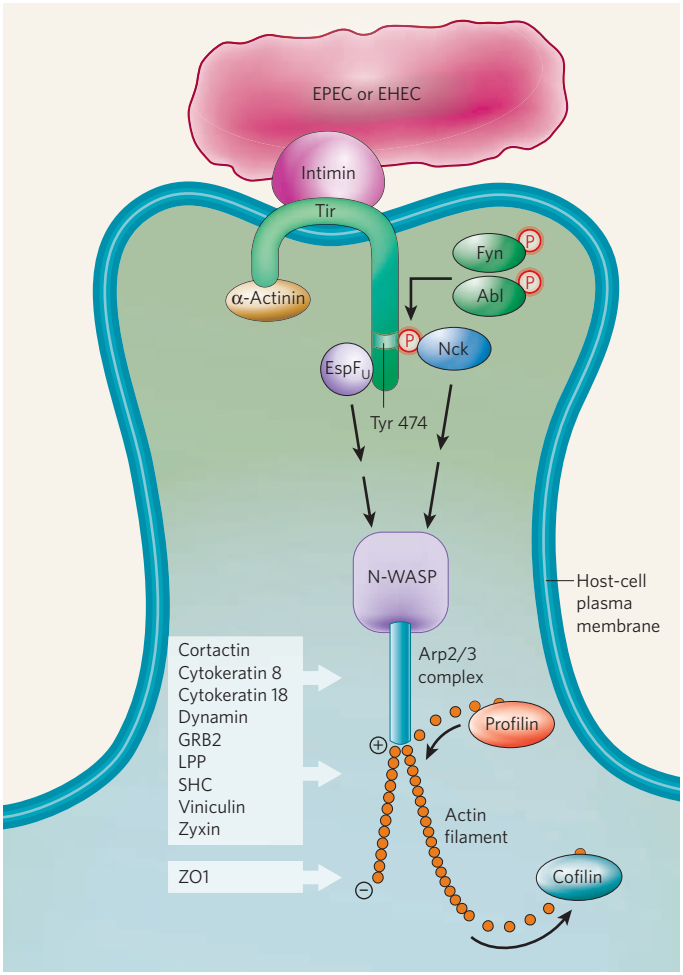


Figure 2 | Generation of pedestals by EPEC and EHEC. During infection with the extracellular bacterium EPEC, the intimin receptor (Tir) translocates into the host cell and inserts itself into the host-cell plasma membrane (a process mediated by the T3SS). This receptor interacts with intimin on the bacterial surface, thereby firmly anchoring the bacterium to the host cell. The carboxy terminus of EPEC Tir becomes phosphorylated on the tyrosine residue at position 474 by at least two host protein kinases, Fyn and Abl, resulting in host adaptor protein Nck being recruited and binding directly to Tir. During infection with EHEC, by contrast, the tyrosine-phosphorylation event is subverted by the EHEC effector EspF$_u$, so Nck is not required. During EPEC or EHEC infection, N-WASP and the Arp2/3 complex (which consists of seven host proteins) are recruited downstream of the Tir-interacting protein (Nck or EspF$_u$), leading to the generation of actin filaments beneath the attached bacteria and the formation of the pedestal structure. Numerous proteins are found in EPEC pedestals (some of which are listed in the shaded box); however, the precise organization of these proteins in EPEC- and EHEC-induced pedestal generation has not been clearly shown. It has been demonstrated that the tight-junction-associated protein ZO1 localizes to the distal portion of the actin filaments of the EPEC pedestal. In addition, the actin-disassembly protein cofilin has been shown to localize to pedestals and presumably, together with the actin-assembly protein profilin, regulates the actin-filament dynamics in pedestals. Also, the amino terminus of Tir has been shown to bind directly to α-actinin, but the effect of this interaction is unknown.

Interactions of bacterial pathogens with signalling pathways

A beneficial strategy used by many pathogens is to interfere with the phosphorylation cascades in the intracellular-signalling pathways of the host cell. Phosphorylation states are usually controlled by protein kinases and protein phosphatases, and the functions of these enzymes are mimicked by certain bacterial effector proteins. Evidence for this comes from the study of T3SS effectors from *S. enterica* and *Yersinia* spp. The *S. enterica* effector SigD (also known as SopB) functions as a phosphoinositide phosphatase that catalyses the dephosphorylation of host phosphatidylinositol-4,5-bisphosphate and phosphatidylinositol-3,4,5-trisphosphate[40]. As a result, membrane-fission dynamics are altered and probably affect SCV formation. The effector YpkA (and its homologue YopO), produced by *Yersinia* spp., has structural and functional similarities to serine/threonine kinases. This effector is secreted by the bacteria in an inactive form and is autophosphorylated, and thereby activated, in the host cell, where it modulates the actin cytoskeleton through a direct interaction with the GTPase RAC1 (ref. 41). An interesting variation of manipulating host intracellular-signalling pathways involves another effector produced by *Yersinia* spp., YopM, which simultaneously binds to (and thereby activates) two host protein kinases[42]. However, the functional significance of the formation of this complex remains poorly understood (Box 2).

Pathogen preservation

It is therefore clear that bacterial pathogens use diverse mechanisms to accomplish a similar goal — to interact with and, potentially, alter the host cell. However, bacterial pathogens must also ensure their own preservation in the host: they need to evade the immune response, and this facilitates replication and spread, which are essential for the success of any pathogen. The following examples highlight the incredibly diverse mechanisms that bacterial pathogens use to evade both innate immune responses and adaptive immune responses.

Inflammation and nuclear factor-κB

A cornerstone of innate immunity is the expression of genes that are responsive to the transcription factor nuclear factor-κB (NF-κB)[43] (Fig. 3). This process is induced after bacterial pathogen-associated molecular patterns (PAMPs) are detected by pattern-recognition receptors (PRRs), including Toll-like receptors and NOD (nucleotide-binding oligomerization-domain protein)-like receptors (NLRs) (see ref. 44 for a review and see page 819). NF-κB-responsive genes include those that encode pro-inflammatory cytokines, anti-apoptotic factors (such as Bcl-2) and defensins (a class of antimicrobial peptide). Before these genes can be transcribed, NF-κB needs to be activated, and this occurs when its cytoplasmic binding partner, inhibitor of NF-κB (IκB), is degraded, enabling NF-κB to translocate to the nucleus. The degradation of IκB occurs after it is phosphorylated by the protein IκB kinase

(IKK), the activity of which is stimulated by PRRs. Phosphorylated IκB is then modified with ubiquitin and undergoes proteolytic degradation (see ref. 45 for a review).

Pathogenic microorganisms have come to 'understand' the NF-κB-activation pathway and have developed strategies to circumvent it (Fig. 3). For example, both *S. flexneri* and *Yersinia* spp. can prevent IκB from being ubiquitylated and therefore prevent its degradation, causing NF-κB to remain inactive in the cell cytoplasm[46]. These bacteria effect this through the T3SS effector proteins OspG (from *S. flexneri*) and YopP/J (YopP and YopJ being orthologous proteins from different species of *Yersinia*). OspG binds to the ubiquitylated form of the E2 ubiquitin-conjugating enzyme UBCH5B (also known as UBE2D2) and prevents the transfer of ubiquitin to IκB by an E3 ubiquitin–protein ligase, even though IκB phosphorylation still occurs[47]. By contrast, until recently, it was known that YopP/J inhibits NF-κB signalling, but it was unclear whether this results from the inhibition of IκB phosphorylation or from the de-ubiquitylation of IκB. Intriguing new biochemical data indicate that inhibition of phosphorylation is the mechanism of action (Box 1).

Interestingly, *S. flexneri* ensures evasion of innate immune responses by altering the NF-κB-activation pathway at several points. Recent work has shown that *S. flexneri* uses the T3SS effector OspF to manipulate the physical and spatial context of DNA encoding NF-κB-responsive genes[48] (Fig. 3). Epigenetic regulation through DNA modifications such as methylation can have marked effects on gene expression[49]. OspF functions as a unique phosphatase (Box 1). It dephosphorylates the mitogen-activated protein kinase (MAPK) ERK2 in the nucleus (Box 2), so ERK2 cannot then activate mitogen- and stress-activated kinase 1 (MSK1) and MSK2 (ref. 50). This, in effect, prevents histone phosphorylation, which is a prerequisite for NF-κB-dependent transcription[48]. Therefore, genes that are usually transcriptionally activated by NF-κB in response to the detection of *S. flexneri* remain silent.

Altering antigen presentation

The adaptive immune response functions together with the innate immune response but is pathogen specific. It therefore initially requires precise pathogen identification, and it culminates in robust induction of immunity. Bacteria are recognized and internalized by specialized cells known as antigen-presenting cells (APCs) (Fig. 4). The microorganisms are then degraded by proteases that are sequestered in specialized membrane-enclosed vesicular compartments. The degraded microbial components (peptides) then bind to host proteins known as major histocompatibility complex (MHC) class II molecules, and these antigen complexes are transported to the host-cell surface, where they are presented to other immune cells.

Several microbial pathogens subvert the initiation of adaptive immune responses. Although the mechanisms that these pathogens use have not been defined completely at the molecular level, it is clear that these mechanisms are diverse. For example, *S. enterica* can block antigen presentation by dendritic cells, an important class of APC, through an incompletely defined mechanism (Fig. 4). This is accomplished by inducing a decrease in the number of peptide-bound MHC class II molecules on the surface of infected dendritic cells[51], resulting in the activation (and proliferation) of fewer T cells, an important class of adaptive immune cell.

Yersinia spp. also alter antigen presentation by dendritic cells (Fig. 4), and the mechanism for this process is better understood than that used by *S. enterica*. *Yersinia enterocolitica* uses the T3SS effector YopP/J, which inhibits adaptive immune responses by affecting the canonical MAPK-signalling pathway (see ref. 52 for a review of MAPK-signalling pathways). YopP/J prevents phosphorylation of the MAPKs JNK and p38 MAPK, leading dendritic cells to take up less antigen by clathrin-mediated endocytosis[53]. Presumably, this interference in the MAPK-signalling cascade results from the acetyltransferase activity of YopP/J, but this has yet to be determined. Nevertheless, by reducing antigen uptake, *Yersinia* spp. elicit a similar response to that induced by *S. enterica*: they restrict the proliferation of T cells and thereby limit the adaptive immune response. These findings that YopP/J has more than one effect raise the intriguing idea that an effector protein can have several functions in the host cell.

Box 2 | Nuclear translocation of bacterial effectors

The *Shigella flexneri* effector OspF is noteworthy not only for its biochemistry but also for its localization. OspF dephosphorylates ERK2 in the host-cell nucleus[48] (Fig. 3). OspF is one of the few bacterial effectors that are known to translocate to the nucleus[48]. Others include the *Salmonella enterica* effector SspH1, the *Yersinia* spp. effector YopM, and another *S. flexneri* effector, IpaH9.8 (Fig. 1e).

Recently, the nuclear localization of these effectors has been considered together with their functions, shedding light on the unique targeting of these effectors. For example, IpaH9.8 lacks a conventional nuclear-localization signal (NLS), but it is targeted to the nucleus[70]. IpaH9.8 has now been shown to bind to the host splicing factor U2AF[35], impairing the U2AF[35]-mediated processing of messenger RNAs that encode products important for pro-inflammatory signalling pathways (for example, interleukin-8)[71]. The interaction of IpaH9.8 and U2AF[35] provides a plausible mechanism for the nuclear translocation of IpaH9.8, given that U2AF[35] shuttles between the cytosol and nucleus.

Similarly, SspH1 (from *S. enterica*) is translocated to the host-cell nucleus, where it decreases the effects of interleukin-8-mediated signalling, by inhibiting NF-κB-dependent gene transcription[72]. The recent finding that the host protein kinase PKN1, which can translocate to the nucleus and decrease NF-κB signalling, binds SspH1 provides a plausible mechanism for both the targeting and the function of this bacterial effector[73]. However, given that PKN1 is a substrate of SspH1 (that is, PKN1 is modified by the E3 ubiquitin–protein ligase activity of SspH1; see Box 1), the interaction between these two proteins is undoubtedly complex.

By contrast, YopM (from *Yersinia* spp.) is targeted to the nucleus by two unconventional NLSs that are present within the coding sequence[74]. Recently, this effector was identified to form a tripartite complex with two host protein kinases: protein-kinase-C-like 2 (PRK2; also known as PKN2) and ribosomal protein S6 kinase 1 (ref. 42). Although this complex could be found in the nucleus, it was much more abundant in the cytosol, and the functional importance of the nuclear targeting of YopM remains in question.

Perhaps the most interesting of the T3SS effectors that translocate to the nucleus are the AvrBs3-like proteins, which are secreted by the plant pathogens *Xanthomonas* spp. These effectors translocate to the nucleus as a result of a carboxy-terminal NLS that is proximal to a transcriptional-activation domain[75]. It is proposed that the AvrBs3-like proteins alter the transcription of host genes directly, although promoters modulated by these proteins have yet to be identified. Effector proteins from mammalian pathogens have not been reported to have a similar activity.

Given the large number of effector proteins that are now being identified to be produced by bacterial pathogens (see the section Coordination of the attack), the existence of multifunctional effectors is noteworthy.

Yersinia spp. and other microbial pathogens have in common another clever strategy to counteract adaptive immune responses. They eradicate APCs by inducing apoptosis. YopP/J is also a key component of this process during infection with *Yersinia* spp. This effect could result from abrogation of the potent anti-apoptotic signalling stimulus provided by NF-κB signalling. For example, NF-κB induces the production of the anti-apoptotic regulator Bcl-2, and, as previously mentioned, YopP/J can disrupt NF-κB signalling effectively. Then, with no APCs present to activate T cells, the proliferation of these immune cells is abrogated, halting the adaptive immune response.

Interestingly, *Shigella* spp. and *S. enterica* induce the death of macrophages, by an unknown NF-κB-independent mechanism that relies on signalling through the key pro-inflammatory activator caspase 1 (refs 7, 54) (Fig. 4). It was thought initially that these bacteria use T3SS effectors to induce cell death. However, recent evidence indicates that caspase 1 activation is mediated through the detection of bacterial flagellin by IPAF (also known as NLRC4), a cytosolic PRR[55]. Intriguingly, the secretion of flagellin into the host-cell cytosol by *S. enterica* depends on a functional T3SS. The mechanism of caspase-1-mediated macrophage death (now termed pyroptosis) is emerging, but the reason

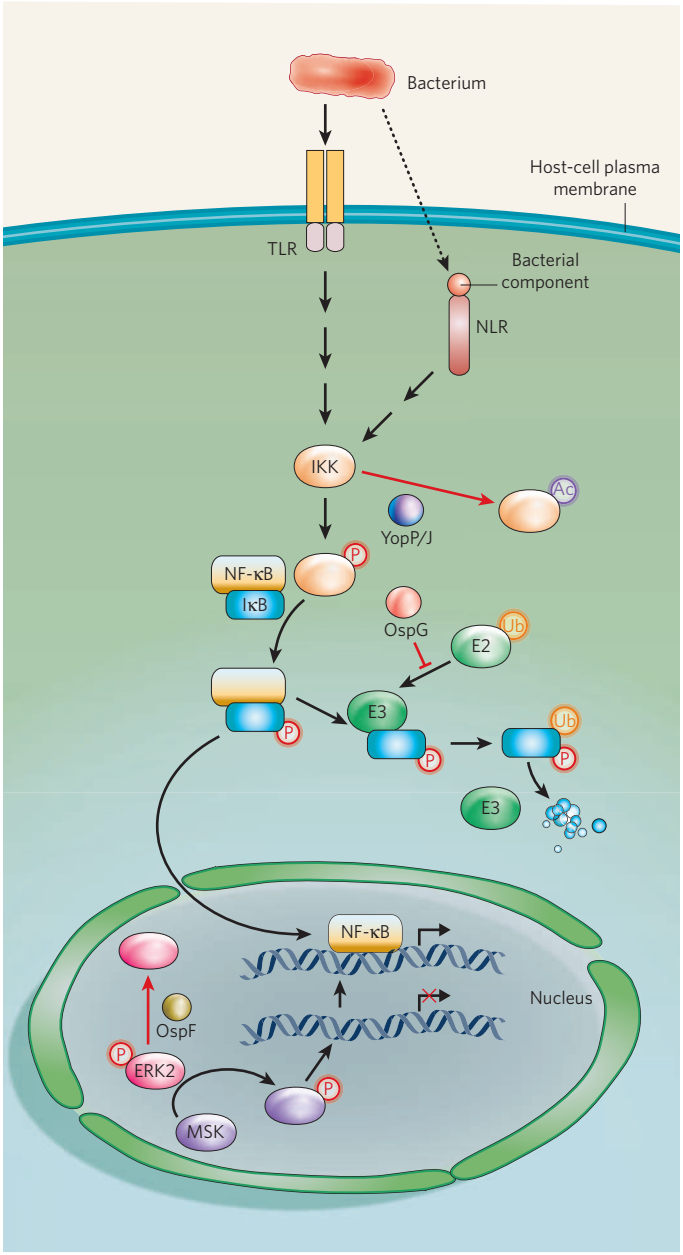


Figure 3 | Subversion of NF-κB-mediated signalling. The transcription factor NF-κB initiates the expression of genes that encode many innate immune factors. The NF-κB-signalling pathway is therefore crucial to the host. Various pathogens can subvert this pathway at different points. After pathogenic bacteria are detected by PRRs (TLRs and/or NLRs), a signalling cascade is triggered, resulting in the phosphorylation of the protein-kinase complex IKK. Activated IKK then catalyses the phosphorylation of the inhibitor of NF-κB, IκB. Ubiquitin, carried by an E2 ubiquitin-conjugating enzyme, is attached to phosphorylated IκB by an E3 ubiquitin–protein ligase, marking IκB for degradation in the cytosol and releasing NF-κB to translocate to the nucleus. The induction of gene expression by NF-κB requires remodelling of chromatin through the phosphorylation of histones. This is mediated by activated MSK, a protein kinase that is activated by the MAPK ERK2. The points at which protein effectors secreted by pathogenic bacteria interfere with NF-κB-mediated gene expression are indicated in red. YopP/J, produced by *Yersinia* spp., disarms IKK by competitively acetylating key amino-acid residues, thereby preventing their phosphorylation. *Shigella flexneri* subverts this signalling pathway at two points, in the cytosol and in the nucleus. The *S. flexneri* effector OspG 'derails' IκB ubiquitylation by binding the ubiquitylated E2 molecule. By contrast, another *S. flexneri* effector, OspF, prevents chromatin remodelling, through dephosphorylating activated ERK2.

for its existence is enigmatic. Indeed, the induction of cell death as a bacterial immune-evasion strategy is controversial. However, the outcome of pathogen-induced host-cell death is context specific and therefore might either harm or benefit the host (see ref. 56 for a review).

The above examples illustrate that bacterial pathogens have evolved different strategies to ensure their preservation when they interact with host cells, and these most commonly involve the manipulation of host-cell intracellular-signalling pathways by secreted bacterial effectors. Considering the coverage of discrete steps targeted in a given signalling pathway, microbial 'understanding' of eukaryotic signalling pathways is truly astounding.

Unresolved issues

Despite progress towards understanding the molecular mechanisms that underlie microbial pathogenesis, much remains to be learned about the bacterial effector proteins that are central to this process. In this section, we outline some of the important unanswered questions about effector acquisition, evolution and coordination.

Selection and evolution of virulence factors

It is clear that microbial pathogens have diverse mechanisms for interacting with, and manipulating, host cells and for evading host immune responses. Presumably, bacterial pathogens have evolved these processes because they provide a selective advantage. Curiously, pathogens often encode several related copies of effectors: for example, EHEC produces at least 14 variants of the T3SS effector NleG, and *L. monocytogenes* produces several internalin proteins. Although, in some cases, these proteins might target related variants of host-cell components, they are presumably redundant in other cases. Why bacterial pathogens maintain such a vast collection of seemingly redundant effectors is controversial, although recent evidence indicates that bacterial effector proteins are also crucial for successful transmission between hosts[57].

It is now understood that the genetic transfer of effector-encoding genes between bacteria (by bacteriophages, conjugation or transformation) has a key role in generating diversity in pathogens (and in generating new diseases). For example, it has recently been shown that a bacterial nucleoid protein, H-NS, can silence genes when they are initially acquired by horizontal transfer. These genes are then integrated into various regulatory pathways, including virulence regulons (in which the silencing is presumably removed), allowing the encoded molecules to participate in virulence pathways[58].

Coordination of the attack

Pathogens with a T3SS commonly have a large repertoire of effectors (often tens to hundreds). For example, recent studies have shown that EHEC has at least 40 T3SS effectors, whereas the plant pathogen *Pseudomonas syringae* has 190 T3SS effectors[59,60]. Similar numbers of effectors are also present in pathogens with a T4SS, such as *Legionella pneumophila*. This raises an important question: how are the expression, secretion and functional delivery of these effectors regulated and, perhaps more crucially, coordinated? Virulence-factor coordination was recently reported in the Gram-positive pathogen *Staphylococcus aureus*: the production of its pore-forming toxin Panton–Valentine leukocidin was shown to increase the production of Spa, a known *S. aureus* virulence factor[61]. The authors of this study suggested that Panton–Valentine leukocidin and Spa function together to exacerbate pathogenesis, although how they do so is unclear. Such effector coordination is also likely to be found in other pathogens with a T3SS or T4SS.

There is no doubt that specialized secretion systems need to coordinate the delivery of effectors to maintain an overall virulence strategy; however, this has not been documented. It also has not been documented, although it is presumed, that adherence precedes effector delivery. In theory, pathogens could deliver different specialized sets of effectors to different types of host cell (for example, a macrophage and an epithelial cell) or to different tissue sites in a host, and different sets of effectors could also be used to target different host species. Moreover, it is conceivable that secretion systems could work in reverse: that is, they could acquire host-cell molecules,

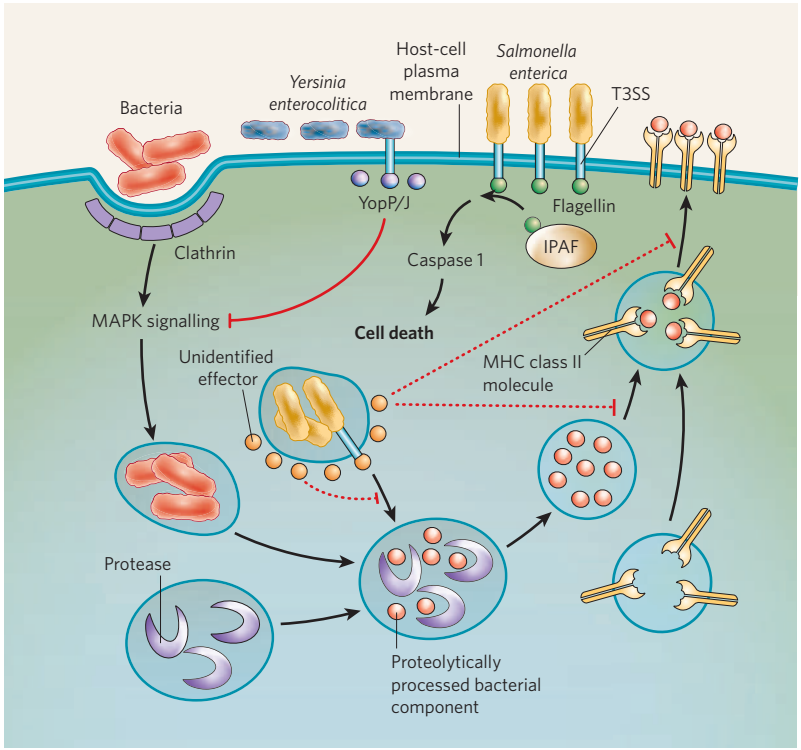


Figure 4 | Prevention of antigen presentation by APCs. Components of the adaptive immune system are activated after they recognize fragments of pathogens that are presented at the surface of infected host cells. Antigens from pathogenic bacteria are presented in this way after the bacteria have been taken up into endocytic vesicles that then fusewith cellular-protease-containing vesicles, in which the bacterial pathogens are degraded. Proteolytically processed bacterial components bind to MHC class II molecules, and the vesicles containing these complexes are transported to the cell surface, where the bacterial peptides are displayed to immune cells. Bacterial pathogens interfere with this antigen processing and presentation pathway at several points, which are indicated in red. *Yersinia enterocolitica* produces YopP/J, which subverts antigen presentation by inhibiting the MAPK-signalling pathway and thereby preventing clathrin-mediated endocytosis. *Salmonella enterica* secretes an unidentified effector that prevents antigen presentation, although the targeted step also remains unidentified (dashed blocking arrows). In addition, *S. enterica* can prevent antigen presentation by inducing cell death. This is thought to occur through the T3SS-mediated delivery of flagellin into the cytosol. Flagellin is subsequently detected by the PRR known as IPAF, which initiates a signalling cascade that results in the activation of caspase 1. Caspase 1 then mediates macrophage death by the newly defined mechanism called pyroptosis. *Shigella* spp. are also thought to induce macrophage death in the same manner.

signals or even energy and nutrients for the bacteria, although there is no evidence to support this idea at present.

A recent bioinformatic study[59] shows that the complexity of these virulence systems might be even greater than has been thought. The authors of this study proposed that pathogenic microorganisms can instantaneously produce novel chimaeric hybrids of T3SS effectors through a process referred to as terminal reassortment[59]. By fusing new protein-coding sequences to sequences that control the expression and secretion of T3SS effectors, microorganisms can 'sample' new combinations of secreted effectors. This seemingly hastened effector evolution, coupled with the propensity of the genes encoding these 'shuffled' effectors to be located on mobile genetic elements and to confer strong selective pressure through either virulence or transmission, suggests that pathogenic microorganisms can coordinate the induced host-cell biology so that pathogenesis is optimized for the benefit of the microorganism. How induced functional responses are coordinated in the host cell is poorly understood.

Pathogenesis in the bigger picture

Pathogenic bacteria have the arduous task of interacting with host cells and reprogramming the complex molecular and cellular networks of these cells to allow bacterial replication and spread, while countering host-defence strategies. Evolution and transmission have shaped this bacterial pursuit through the accumulation of (sometimes) vast arsenals of genes that encode effector proteins, which are probably subject to complex regulation. Piecemeal study of these arsenals might help to define them, but a deeper understanding of their mechanisms and of potential intervention points would best be achieved by considering the delivery, coordination and mechanistic functions of these arsenals as a whole. The current challenge is to assemble a cross-disciplinary toolbox that will enable pathogenesis to be studied at the 'systems' level.

Conclusions

The field of bacterial pathogenesis is a rapidly evolving and expanding one. As the vastness of effector functions is being realized, it is a considerable challenge to integrate the numerous host-cell targets and to translate this knowledge into an accurate understanding of the mechanisms by which effector proteins cause disease.

Moving from studying cultured cells to relevant animal disease models is crucial for understanding disease, yet such studies are often neglected, because cell-culture-based systems are easier to manipulate. However, the opportunity to study pathogenesis in relevant animal models is now within reach, because the current understanding of the mechanistic details of the host–pathogen interface, some of which have been outlined in this article, allows a directed approach to the problem. An elegant example is the recent re-engineering of an *L. monocytogenes* internalin protein to extend the host range of this bacterium to include mice[62]. Conversely, the host can now be engineered such that it is susceptible to infection[63]. These two studies present a glorious opportunity to probe the host–pathogen interface during disease. Similarly, genomic studies have led to the identification of mutations in humans that alter the outcome of bacterial infections[64]. The recent realization that the host microbiota has a crucial role in mediating the outcome of disease adds another layer of complexity[65].

Recognizing that pathogens can overrun crucial host-cell pathways by using a myriad of mechanisms has led to an increased understanding of microbiology, cell biology, biochemistry and immunology. However, this knowledge now needs to be advanced to the point at which it can be translated into a true understanding of disease. This remains the crucial challenge to all who are involved in this field. Only then will it be possible to target these effector mechanisms rationally as a preventive or therapeutic strategy. ■

1. Galan, J. E. & Wolf-Watz, H. Protein delivery into eukaryotic cells by type III secretion machines. *Nature* **444,** 567–573 (2006).
2. Pizarro-Cerda, J. & Cossart, P. Bacterial adhesion and entry into host cells. *Cell* **124,** 715–727 (2006).
3. Cossart, P. & Sansonetti, P. J. Bacterial invasion: the paradigms of enteroinvasive pathogens. *Science* **304,** 242–248 (2004).
4. Stevens, J. M., Galyov, E. E. & Stevens, M. P. Actin-dependent movement of bacterial pathogens. *Nature Rev. Microbiol.* **4,** 91–101 (2006).
5. Finlay, B. B. Bacterial virulence strategies that utilize Rho GTPases. *Curr. Top. Microbiol. Immunol.* **291,** 1–10 (2005).
6. Meresse, S. *et al.* Controlling the maturation of pathogen-containing vacuoles: a matter of life and death. *Nature Cell Biol.* **1,** E183–E188 (1999).
7. Gao, L. & Abu Kwaik, Y. Hijacking of apoptotic pathways by bacterial pathogens. *Microbes Infect.* **2,** 1705–1719 (2000).
8. Finlay, B. B. & McFadden, G. Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. *Cell* **124,** 767–782 (2006).
9. Hardt, W. D., Chen, L. M., Schuebel, K. E., Bustelo, X. R. & Galan, J. E. *S. typhimurium* encodes an activator of Rho GTPases that induces membrane ruffling and nuclear responses in host cells. *Cell* **93,** 815–826 (1998).
10. Stender, S. *et al.* Identification of SopE2 from *Salmonella typhimurium*, a conserved guanine nucleotide exchange factor for Cdc42 of the host cell. *Mol. Microbiol.* **36,** 1206–1221 (2000).

11. Zhou, D., Chen, L. M., Hernandez, L., Shears, S. B. & Galan, J. E. A *Salmonella* inositol polyphosphatase acts in conjunction with other bacterial effectors to promote host cell actin cytoskeleton rearrangements and bacterial internalization. *Mol. Microbiol.* **39,** 248–259 (2001).
12. Alto, N. M. *et al.* Identification of a bacterial type III effector family with G protein mimicry functions. *Cell* **124,** 133–145 (2006).
13. Egile, C. *et al.* Activation of the CDC42 effector N-WASP by the *Shigella flexneri* IcsA protein promotes actin nucleation by Arp2/3 complex and bacterial actin-based motility. *J. Cell Biol.* **146,** 1319–1332 (1999).
14. Chakraborty, T. *et al.* A focal adhesion factor directly linking intracellularly motile *Listeria monocytogenes* and *Listeria ivanovii* to the actin-based cytoskeleton of mammalian cells. *EMBO J.* **14,** 1314–1321 (1995).
15. Welch, M. D., Iwamatsu, A. & Mitchison, T. J. Actin polymerization is induced by Arp2/3 protein complex at the surface of *Listeria monocytogenes*. *Nature* **385,** 265–269 (1997).
16. Gruenheid, S. *et al.* Enteropathogenic *E. coli* Tir binds Nck to initiate actin pedestal formation in host cells. *Nature Cell Biol.* **3,** 856–859 (2001).
17. Campellone, K. G., Robbins, D. & Leong, J. M. $EspF_U$ is a translocated EHEC effector that interacts with Tir and N-WASP and promotes Nck-independent actin assembly. *Dev. Cell* **7,** 217–228 (2004).
18. Garmendia, J. *et al.* TccP is an enterohaemorrhagic *Escherichia coli* O157:H7 type III effector protein that couples Tir to the actin-cytoskeleton. *Cell. Microbiol.* **6,** 1167–1183 (2004).
19. Shaner, N. C., Sanger, J. W. & Sanger, J. M. Actin and α-actinin dynamics in the adhesion and motility of EPEC and EHEC on host cells. *Cell Motil. Cytoskeleton* **60,** 104–120 (2005).
20. Cantarelli, V. V. *et al.* Cortactin is necessary for F-actin accumulation in pedestal structures induced by enteropathogenic *Escherichia coli* infection. *Infect. Immun.* **70,** 2206–2209 (2002).
21. Goosney, D. L., DeVinney, R. & Finlay, B. B. Recruitment of cytoskeletal and signaling proteins to enteropathogenic and enterohemorrhagic *Escherichia coli* pedestals. *Infect. Immun.* **69,** 3315–3322 (2001).
22. Goosney, D. L. *et al.* Enteropathogenic *E. coli* translocated intimin receptor, Tir, interacts directly with α-actinin. *Curr. Biol.* **10,** 735–738 (2000).
23. Unsworth, K. E. *et al.* Dynamin is required for F-actin assembly and pedestal formation by enteropathogenic *Escherichia coli* (EPEC). *Cell. Microbiol.* **9,** 438–449 (2007).
24. Batchelor, M. *et al.* Involvement of the intermediate filament protein cytokeratin-18 in actin pedestal formation during EPEC infection. *EMBO Rep.* **5,** 104–110 (2004).
25. Hanajima-Ozawa, M. *et al.* Enteropathogenic *Escherichia coli*, *Shigella flexneri*, and *Listeria monocytogenes* recruit a junctional protein, zonula occludens-1, to actin tails and pedestals. *Infect. Immun.* **75,** 565–573 (2007).
26. Yoshida, S. *et al. Shigella* deliver an effector protein to trigger host microtubule destabilization, which promotes Rac1 activity and efficient bacterial internalization. *EMBO J.* **21,** 2923–2935 (2002).
27. Hardwidge, P. R. *et al.* Modulation of host cytoskeleton function by the enteropathogenic *Escherichia coli* and *Citrobacter rodentium* effector protein EspG. *Infect. Immun.* **73,** 2586–2594 (2005).
28. Hu, L. & Kopecko, D. J. *Campylobacter jejuni* 81-176 associates with microtubules and dynein during invasion of human intestinal cells. *Infect. Immun.* **67,** 4171–4182 (1999).
29. Roy, C. R. & Tilney, L. G. The road less traveled: transport of *Legionella* to the endoplasmic reticulum. *J. Cell Biol.* **158,** 415–419 (2002).
30. Knodler, L. A. & Steele-Mortimer, O. Taking possession: biogenesis of the *Salmonella*-containing vacuole. *Traffic* **4,** 587–599 (2003).
31. Nagai, H. *et al.* A C-terminal translocation signal required for Dot/Icm-dependent delivery of the *Legionella* RalF protein to host cells. *Proc. Natl Acad. Sci. USA* **102,** 826–831 (2005).
32. Robinson, C. G. & Roy, C. R. Attachment and fusion of endoplasmic reticulum with vacuoles containing *Legionella pneumophila*. *Cell. Microbiol.* **8,** 793–805 (2006).
33. Murata, T. *et al.* The *Legionella pneumophila* effector protein DrrA is a Rab1 guanine nucleotide-exchange factor. *Nature Cell Biol.* **8,** 971–977 (2006).
34. Steele-Mortimer, O., Meresse, S., Gorvel, J. P., Toh, B. H. & Finlay, B. B. Biogenesis of *Salmonella typhimurium*-containing vacuoles in epithelial cells involves interactions with the early endocytic pathway. *Cell. Microbiol.* **1,** 33–49 (1999).
35. Cuellar-Mata, P. *et al.* Nramp1 modifies the fusion of *Salmonella typhimurium*-containing vacuoles with cellular endomembranes in macrophages. *J. Biol. Chem.* **277,** 2258–2265 (2002).
36. Drecktrah, D., Knodler, L. A., Howe, D. & Steele-Mortimer, O. *Salmonella* trafficking is defined by continuous dynamic interactions with the endolysosomal system. *Traffic* **8,** 212–225 (2007).
37. Sansonetti, P. J., Ryter, A., Clerc, P., Maurelli, A. T. & Mounier, J. Multiplication of *Shigella flexneri* within HeLa cells: lysis of the phagocytic vacuole and plasmid-mediated contact hemolysis. *Infect. Immun.* **51,** 461–469 (1986).
38. Veiga, E. & Cossart, P. *Listeria* hijacks the clathrin-dependent endocytic machinery to invade mammalian cells. *Nature Cell Biol.* **7,** 894–900 (2005).
39. Shaughnessy, L. M., Hoppe, A. D., Christensen, K. A. & Swanson, J. A. Membrane perforations inhibit lysosome fusion by altering pH and calcium in *Listeria monocytogenes* vacuoles. *Cell. Microbiol.* **8,** 781–792 (2006).
40. Terebiznik, M. R. *et al.* Elimination of host cell $PtdIns(4,5)P_2$ by bacterial SigD promotes membrane fission during invasion by *Salmonella*. *Nature Cell Biol.* **4,** 766–773 (2002).
41. Prehna, G., Ivanov, M. I., Bliska, J. B. & Stebbins, C. E. *Yersinia* virulence depends on mimicry of host Rho-family nucleotide dissociation inhibitors. *Cell* **126,** 869–880 (2006).
42. McDonald, C., Vacratsis, P. O., Bliska, J. B. & Dixon, J. E. The *Yersinia* virulence factor YopM forms a novel protein complex with two cellular kinases. *J. Biol. Chem.* **278,** 18514–18523 (2003).
43. Hayden, M. S., West, A. P. & Ghosh, S. NF-κB and the immune response. *Oncogene* **25,** 6758–6780 (2006).
44. Akira, S., Uematsu, S. & Takeuchi, O. Pathogen recognition and innate immunity. *Cell* **124,** 783–801 (2006).
45. Perkins, N. D. Post-translational modifications regulating the activity and function of the nuclear factor κB pathway. *Oncogene* **25,** 6717–6730 (2006).
46. Angot, A., Vergunst, A., Genin, S. & Peeters, N. Exploitation of eukaryotic ubiquitin signaling pathways by effectors translocated by bacterial type III and type IV secretion systems. *PLoS Pathog.* **3,** e3 (2007).
47. Kim, D. W. *et al.* The *Shigella flexneri* effector OspG interferes with innate immune responses by targeting ubiquitin-conjugating enzymes. *Proc. Natl Acad. Sci. USA* **102,** 14046–14051 (2005).
48. Arbibe, L. *et al.* An injected bacterial effector targets chromatin access for transcription factor NF-κB to alter transcription of host genes involved in immune responses. *Nature Immunol.* **8,** 47–56 (2007).
49. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
50. Soloaga, A. *et al.* MSK2 and MSK1 mediate the mitogen- and stress-induced phosphorylation of histone H3 and HMG-14. *EMBO J.* **22,** 2788–2797 (2003).
51. Cheminay, C., Mohlenbrink, A. & Hensel, M. Intracellular *Salmonella* inhibit antigen presentation by dendritic cells. *J. Immunol.* **174,** 2892–2899 (2005).
52. Ashwell, J. D. The many paths to p38 mitogen-activated protein kinase activation in the immune system. *Nature Rev. Immunol.* **6,** 532–540 (2006).
53. Autenrieth, S. E. *et al. Yersinia enterocolitica* YopP inhibits MAP kinase-mediated antigen uptake in dendritic cells. *Cell. Microbiol.* **9,** 425–437 (2007).
54. Scott, A. M. & Saleh, M. The inflammatory caspases: guardians against infections and sepsis. *Cell Death Differ.* **14,** 23–31 (2007).
55. Miao, E. A. *et al.* Cytoplasmic flagellin activates caspase-1 and secretion of interleukin-1β via Ipaf. *Nature Immunol.* **7,** 569–575 (2006).
56. DeLeo, F. R. Modulation of phagocyte apoptosis by bacterial pathogens. *Apoptosis* **9,** 399–413 (2004).
57. Wickham, M. E., Brown, N. F., Boyle, E. C., Coombes, B. K. & Finlay, B. B. Virulence is positively selected by transmission success between mammalian hosts. *Curr. Biol.* **17,** 783–788 (2007).
58. Navarre, W. W. *et al.* Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science* **313,** 236–238 (2006).
59. Stavrinides, J., Ma, W. & Guttman, D. S. Terminal reassortment drives the quantum evolution of type III effectors in bacterial pathogens. *PLoS Pathog.* **2,** e104 (2006).
60. Tobe, T. *et al.* An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc. Natl Acad. Sci. USA* **103,** 14941–14946 (2006).
61. Labandeira-Rey, M. *et al. Staphylococcus aureus* Panton–Valentine leukocidin causes necrotizing pneumonia. *Science* **315,** 1130–1133 (2007).
62. Wollert, T. *et al.* Extending the host range of *Listeria monocytogenes* by rational protein design. *Cell* **129,** 891–902 (2007).
63. Lecuit, M. *et al.* A transgenic model for listeriosis: role of internalin in crossing the intestinal barrier. *Science* **292,** 1722–1725 (2001).
64. Casanova, J. L. & Abel, L. Human genetics of infectious diseases: a unified theory. *EMBO J.* **26,** 915–922 (2007).
65. Lupp, C. *et al.* Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae. *Cell Host Microbe* **2,** 119–129 (2007).
66. Rohde, J. R., Breitkreutz, A., Chenal, A., Sansonetti, P. J. & Parsot, C. Type III secretion effectors of the IpaH family are E3 ubiquitin ligases. *Cell Host Microbe* **1,** 77–83 (2007).
67. Rytkonen, A. *et al.* SseL, a *Salmonella* deubiquitinase required for macrophage killing and virulence. *Proc. Natl Acad. Sci. USA* **104,** 3502–3507 (2007).
68. Mukherjee, S. *et al. Yersinia* YopJ acetylates and inhibits kinase activation by blocking phosphorylation. *Science* **312,** 1211-1214 (2006).
69. Li, H. *et al.* The phosphothreonine lyase activity of a bacterial type III effector family. *Science* **315,** 1000–1003 (2007).
70. Toyotome, T. *et al. Shigella* protein $IpaH_{9.8}$ is secreted from bacteria within mammalian cells and transported to the nucleus. *J. Biol. Chem.* **276,** 32071–32079 (2001).
71. Okuda, J. *et al. Shigella* effector IpaH9.8 binds to a splicing factor $U2AF^{35}$ to modulate host immune responses. *Biochem. Biophys. Res. Commun.* **333,** 531–539 (2005).
72. Haraga, A. & Miller, S. I. A *Salmonella enterica* serovar Typhimurium translocated leucine-rich repeat effector protein inhibits NF-κB-dependent gene expression. *Infect. Immun.* **71,** 4052–4058 (2003).
73. Haraga, A. & Miller, S. I. A *Salmonella* type III secretion effector interacts with the mammalian serine/threonine protein kinase PKN1. *Cell. Microbiol.* **8,** 837–846 (2006).
74. Benabdillah, R., Mota, L. J., Lutzelschwab, S., Demoinet, E. & Cornelis, G. R. Identification of a nuclear targeting signal in YopM from *Yersinia* spp. *Microb. Pathog.* **36,** 247–261 (2004).
75. Schornack, S., Meyer, A., Romer, P., Jordan, T. & Lahaye, T. Gene-for-gene-mediated recognition of nuclear-targeted AvrBs3-like bacterial effector proteins. *J. Plant Physiol.* **163,** 256–272 (2006).

Acknowledgements We thank members of B.B.F.'s laboratory for helpful discussions and critical reading of the manuscript. We gratefully acknowledge F. Ness for assistance with the preparation of figures. We apologize to authors whose work could not be cited as a result of space restrictions. Work in B.B.F.'s laboratory is supported by grants from the Canadian Institutes of Health Research (CIHR), the Howard Hughes Medical Institute (HHMI), the Foundation for the National Institutes of Health, and Genome Canada. A.P.B. is supported by fellowships from the CIHR and the Michael Smith Foundation for Health Research (MSFHR). J.A.G. is supported by a Canadian Association for Gastroenterology/CIHR/AstraZeneca fellowship and a fellowship from the MSFHR. B.B.F. is a CIHR Distinguished Investigator, an HHMI International Research Scholar, and the Peter Wall Distinguished Professor, at the University of British Columbia.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to B.B.F. (bfinlay@interchange.ubc.ca).

Bacterial pathogenomics

Mark J. Pallen[1] & Brendan W. Wren[2]

Genomes from all of the crucial bacterial pathogens of humans, plants and animals have now been sequenced, as have genomes from many of the important commensal, symbiotic and environmental microorganisms. Analysis of these sequences has revealed the forces that shape pathogen evolution and has brought to light unexpected aspects of pathogen biology. The finding that horizontal gene transfer and genome decay have key roles in the evolution of bacterial pathogens was particularly surprising. It has also become evident that even the definitions for 'pathogen' and 'virulence factor' need to be re-evaluated.

The sequencing of bacterial genomes (see Glossary) has occurred against the backdrop of an established programme of research on bacterial pathogenesis. Nonetheless, it has uncovered aspects of pathogen biology that were unexpected before the genomic revolution. Here, we examine the 'creative clash' between genomic research and bacterial pathogenesis research, an encounter that has spawned new technologies and new avenues for applied research. In addition, we discuss the forces that have shaped the evolution of bacterial pathogens, and we reappraise human–pathogen interactions in the light of bacterial ecology and evolution.

Genome dynamics

At the start of the genomic era, each project to sequence a bacterial genome was viewed as equivalent in difficulty to a Moon landing. However, the cutting edge soon shifted to determining the genome sequences of multiple strains in each species[1–4], heralding a transformation in our view of bacterial diversity. Comparisons between the genomes of related strains and species of bacterial pathogens, across the whole range of taxonomic variation, have made it clear that a 'one size fits all' approach cannot be applied to the evolutionary dynamics of bacterial virulence[5–7]. Instead, different evolutionary processes predominate in different taxonomic groups.

Three main forces have been found to shape genome evolution: gene gain, gene loss and gene change (that is, any changes that affect the sequences or order of the existing genes) (Fig. 1). In the genome of some bacterial pathogens (for example, *Yersinia pestis*[8]), all three are evident. In addition, differences in the scale and the timing of these changes in different lineages of bacterial pathogens have resulted in at least three main patterns of genome dynamics. First, some genetically uniform lineages, which are also usually reproductively isolated, have emerged recently in evolutionary terms (for example, *Bacillus anthracis* and *Mycobacterium leprae*). Second, recombination can occur between closely related sequences in closely related strains; this is common in naturally competent mucosal pathogens (for example, *Neisseria meningitidis*, *Haemophilus influenzae* and *Streptococcus pneumoniae*). Third, widespread horizontal gene transfer, bringing in new sequences, predominates in certain pathogens (for example, many enterobacteria, and some staphylococci and streptococci).

The smallest-scale variation in bacterial genomes occurs at the level of single-nucleotide polymorphisms (SNPs). SNP detection has been applied extensively to recently emerged genetically uniform pathogens, such as *M. leprae*, *Mycobacterium tuberculosis*, *Y. pestis and B. anthracis* (the last driven by forensic considerations after the anthrax attacks of 2001 in the United States)[9–12]. More recently, whole-genome sequencing has been used to detect SNPs in more variable species, such as *Escherichia coli* and *Francisella tularensis*[13–15]. This approach to SNP detection enabled *E. coli* strains that had diverged for as few as 200 generations to be differentiated[16] and revealed genomic changes in *Burkholderia mallei* after accidental human infection[17]. These studies indicate that the use of whole-genome sequencing could soon become a routine epidemiological tool in bacteriology, as it already is in virology (Box 1).

Genome sequencing has also confirmed that phase variation is a widespread source of intraspecific genotypic and phenotypic variation[18,19]. Several mutational mechanisms are exploited by bacteria to switch gene and/or protein expression on or off. For example, in *Campylobacter jejuni*, the presence of several tens of homopolymeric nucleotide repeat sequences can lead to slippage during DNA replication, resulting in a varied repertoire of structures exposed on the bacterial-cell surface[20]. By contrast, *Bacteroides fragilis* uses DNA inversion to modulate more than 20 genetic loci, which contain genes that encode bacterial surface proteins, polysaccharides and components of regulatory systems[21]. The combinatorial mathematics of phase variation mean that a bacterium with just 20 phase-variable loci can exist in 2^{20} (that is, more than a million) different states.

Horizontal gene transfer

The greatest surprise resulting from the application of genomics to bacteriology was the extent of genomic variability within many bacterial species. Two *E. coli* strains can differ by as much as a quarter

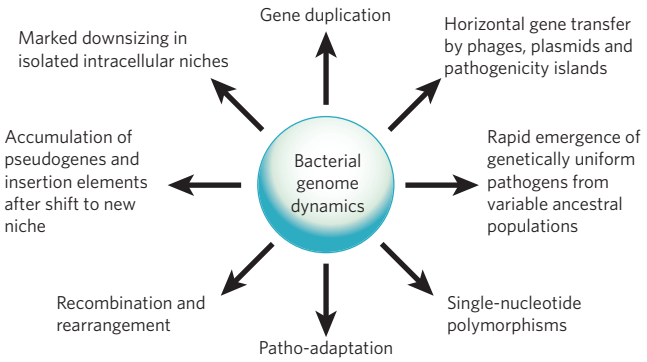


Figure 1 | Bacterial genome dynamics. There are three main forces that shape bacterial genomes: gene gain, gene loss and gene change. All three of these can take place in a single bacterium. Some of the changes that result from the interplay of these forces are shown.

[1]Centre for Systems Biology, University of Birmingham, Birmingham B15 2TT, UK. [2]Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

Box 1 | Technologies and applications in pathogenomics

In 1995, J. Craig Venter and colleagues showed that genomes could be sequenced efficiently by using a whole-genome shotgun approach, facilitated by computational sequence assembly and laboratory-based finishing and closure approaches[85]. This is now the conventional approach in almost all genome-sequencing projects. Arguments continue about whether it is more efficient and informative to obtain complete genome sequences for a small number of strains or to obtain partial sequences for a large collection of strains[86]. Highly efficient and cost-effective, 'next-generation', sequencing technologies hold promise for sequencing many more bacterial genomes at a cost of only a few hundred dollars each[61].

In the field of bacteriology, the genomic revolution catalysed the development of bioinformatics approaches (for example, for comparing genomes and for detecting sequence-based evidence of selection and horizontal gene transfer) and high-throughput experimental technologies (for example, microarray-based transcriptomics, mass mutagenesis, and the use of simpler surrogate hosts such as the microscopic nematode *Caenorhabditis elegans* and the budding yeast *Saccharomyces cerevisiae*[87,88]). The advent of genomics also led to new experimental approaches for assessing genomic variability, including multi-locus sequence typing[89], variable-number tandem-repeat typing[90] and the use of microarrays for genomic comparisons[91]. These approaches have provided insight into the population structure of bacterial pathogens, facilitated the classification of strains by source or pathogenicity and helped to identify pathogen-specific genes in pathogenic lineages[92].

Practical applications of sequencing bacterial genomes include metabolic reconstruction (allowing the design of improved culture media[93]), glyco-engineering (allowing the biotechnological manipulation of sugar residues on macromolecules)[94] and reverse vaccinology (facilitating the discovery of new vaccine targets[95]). Genomic studies have also focused on the host genome, allowing the identification of host genes that are expressed after bacterial invasion[96] or are associated with susceptibility to infection[97].

of their genomes: for example, the laboratory strain *E. coli* K-12 is missing 1.4 megabases of DNA present in *E. coli* O157 (ref. 3). For many important pathogens, the genes common to all strains within a species (known as the core genome) are a minority component of the entire gene pool for that species (the pan-genome). Furthermore, a distinction can be made between closed pan-genomes and open pan-genomes. For closed pan-genomes, completing the genome sequencing of additional bacterial strains is unlikely to yield new genes. By contrast, for open pan-genomes, each new genome sequence reveals new members of the gene pool for that species[22].

The genomes of some bacterial pathogens have gained genes through gene duplication, resulting in increased numbers of key gene clusters or the expansion of important protein families: for example, in *M. tuberculosis*, the gene families encoding acidic glycine-rich proteins and the gene clusters encoding the secreted protein ESAT6 (early secretory antigenic target 6) and its homologues have undergone extensive rounds of gene duplication[23]. Nonetheless, gene gain as a result of horizontal gene transfer remains the most potent source of 'innovation' and variation. However, unlike viruses, bacteria seldom acquire 'eukaryotic-like' genes from their hosts (although there seem to be some exceptions, for example, *Legionella pneumophila*[24]). Instead, horizontal gene transfer generally occurs between different strains and species of bacteria.

Horizontal gene transfer is mediated by diverse mobile genetic elements, including plasmids, bacteriophages and pathogenicity islands (Table 1). These elements often carry genes that encode factors involved in infection (often termed virulence factors) (Box 2). For example, numerous virulence factors and systems are encoded on plasmids. These virulence-associated plasmids can be large (for example, the plant pathogen *Ralstonia solanacearum* carries such a plasmid of more than 2 megabases[25]). They can also be promiscuous: that is, they can move freely between cells with markedly different chromosomal backgrounds. In the extreme case of enterotoxigenic *E. coli*, the association of promiscuous plasmids with diverse chromosomal lineages is all that defines the pathotype of these bacteria[26].

Bacteriophages (that is, bacterial viruses) can also mediate horizontal gene transfer. Some classic virulence factors, such as diphtheria toxin, are encoded in the genomes of bacteriophages that have integrated into the bacterial chromosome (which are known as prophages)[27]. Genomic analyses show that prophages have a widespread role in driving the diversification of bacterial pathogens as distinct as *E. coli*, *Streptococcus*

Table 1 | Examples of mobile genetic elements that encode virulence factors and are present in human pathogens

Type of mobile element	Pathogen	Virulence factor
Plasmid	*Bacillus anthracis*	Anthrax toxin
	Clostridium tetani	Tetanus toxin
	Enterotoxigenic *Escherichia coli*	Heat-stabile toxin, heat-labile toxin and fimbriae
	Mycobacterium ulcerans	Polyketide toxin
	Salmonella enterica serovar Typhimurium	SpvR, SpvA, SpvB, SpvC and SpvD proteins*
	Shigella spp.	Type III secretion system
	Staphylococcus aureus	Exfoliatin B
	Pathogenic *Yersinia* spp.	Type III secretion system
Prophage	*Corynebacterium diphtheriae*	Diphtheria toxin
	Enterohaemorrhagic *E. coli*	Shiga toxin and type III secretion effectors
	S. aureus	Staphylococcal enterotoxin A, exfoliatin A and Panton–Valentine leukocidin
	Streptococcus pyogenes	Streptococcal pyrogenic exotoxins, DNases and streptococcal phospholipase A_2 (Sla)
	Vibrio cholerae	Cholera toxin
Pathogenicity island	*Clostridium difficile*	Clostridial enterotoxin and clostridial cytotoxin
	Enteropathogenic and enterohaemorrhagic *E. coli*	Type III secretion system
	Uropathogenic *E. coli*	Fimbriae, iron-uptake systems, the capsular polysaccharide and α-haemolysin
	Helicobacter pylori	Cag antigen
	S. enterica	Type III secretion systems
	S. aureus	Toxic-shock toxin, staphylococcal enterotoxin B, enterotoxin C, enterotoxin K and enterotoxin L

*Involved in intracellular survival.

pyogenes and *Staphylococcus aureus*[28–30]. Prophages, particularly those derived from tailed bacteriophages, often carry genes that are superfluous for bacteriophage replication, and these genes are present within distinct 'passenger compartments' at one end of the prophage genome. These compartments are sometimes called morons to reflect that the associated prophage genomes encode more DNA than is necessary for bacteriophage replication alone[31]. The genes in these compartments are often implicated in virulence and can show a bias in base composition that sets them apart from the rest of the prophage and from the genome of the bacterial host. For example, in *E. coli* O157, the passenger compartments of lambdoid prophages contain genes with a low G+C composition that encode effector proteins capable of translocation into host cells by a type III secretion mechanism[29].

Pathogenicity islands are another class of mobile element involved in horizontal gene transfer. The term 'pathogenicity island' originated from the study of uropathogenic *E. coli* but has subsequently been widely applied to bacterial pathogens[32]. Pathogenicity islands are usually defined by five characteristics. First, they are clusters of contiguous genes that are present in some related strains or species but not in others. Second, they are presumed to have been acquired by horizontal gene transfer. Third, they are generally associated with transfer RNA gene loci. Fourth, they typically have a G+C content that differs from that of the host bacterial genome. Fifth, they confer on the host bacterium a complex and distinctive virulence phenotype in a single step. Although some pathogenicity islands carry genes encoding integrases (enzymes that integrate the pathogenicity island into the host DNA), the mechanisms underlying the transfer of pathogenicity islands from one genome to another are unclear in many cases, as is the identity of the donor microorganisms.

Despite their mobility, pathogenicity islands are remarkably well integrated into the global regulatory network of bacterial cells. For example, numerous external factors affect the expression of genes on the locus of enterocyte effacement (LEE) pathogenicity island of *E. coli*, making it part of the 'genomic continent'[33–35]. It is also important to recognize that some 'pathogenicity islands' are deletions in one lineage rather than insertions in another[36]. Therefore, instead of considering the evolution of pathogens as a series of acquisitions of pathogenicity islands, a more sophisticated outlook is that genomes are 'molecular palimpsests': that is, the variable compartment of the genome bears the scars of repeated rounds of gene acquisition and erosion.

Gene loss

Bacterial genomes remain about the same size despite the pervasive effects of horizontal gene transfer, so gene gain must be balanced by gene loss[37]. Indeed, it is expected that any gene that is not maintained by natural selection is lost: bacterial genomes are subject to the 'use it or lose it' maxim. Genome sequencing has now provided a series of 'snapshots' that show directly the dynamic processes of gene loss and genome decay (that is, the progressive purging from the genome of unnecessary genes). For example, in many *E. coli* lineages, the Flag-2 and ETT2 gene clusters — which, when intact, span tens of kilobases — have been reduced to small scars occupying only a few hundred base pairs, presumably because they no longer provide any selective advantage to the organism[36,38].

The most surprising snapshots of genome decay have come from recently emerged pathogens that have changed lifestyle, usually to live in a simpler host-associated niche. For example, the genomes of *M. leprae*[39], *Y. pestis*[40] and *Salmonella enterica* serovar Typhi[41] contain hundreds or even thousands of pseudogenes; in the *M. leprae* genome, there are nearly as many pseudogenes as functional genes[39]. These examples contradict the view that every gene in a bacterial genome must have a function and that bacterial genomes never contain 'junk' DNA. Instead, every genome should be viewed as a work in progress, burdened with some non-functional 'baggage of history'.

Another common feature of recently emerged genetically uniform pathogens is the 'proliferation' of transposable elements, particularly insertion sequences, in the genome[42]. This abundance of insertion sequences facilitates homologous recombination within the genome, a process that can result in large-scale chromosomal rearrangements that disrupt the ancestral gene order. In the case of *Y. pestis*, recombination between insertion sequences results in marked anomalies in GC skew (usually a marker of the direction of replication for a given region of chromosome) and in reversible chromosomal rearrangements during *in vitro* growth of the organism[40]. It is unclear whether such large-scale genomic rearrangements have functional relevance.

The most extreme form of genome decay is seen in host-associated bacteria, particularly endosymbionts that have been isolated for long periods in a static and 'undemanding' intracellular niche[43]. Pioneering studies by Siv Andersson and colleagues established that certain intracellular bacteria, such as *Rickettsia prowazekii*, have undergone considerable genomic downsizing, shedding many (or even most) of their ancestral genes[44]. *Buchnera aphidicola*, an aphid endosymbiont, is a pertinent example in that it is a close relative of *E. coli* but has fewer than one-tenth of the genes present in the latter[45]. Extreme genome decay is often accompanied by a shift towards a low G+C content: the largest known shift is in the 160-kilobase genome of the psyllid (jumping plant lice) symbiont *Carsonella ruddii*, which has a G+C content of only 16.5% (ref. 46). But perhaps the most extreme example of bacterial

Box 2 | Defining virulence

In the late nineteenth century, Robert Koch laid the groundwork for establishing a link between pathogens and disease, by putting forward what are now known as Koch's postulates. These postulates are four criteria for determining that a particular organism is the causative agent of a particular disease. First, the organism should be detected in all individuals suffering from the disease but not in their healthy counterparts. Second, it must be possible to isolate the organism from a diseased individual. Third, it must be possible to grow the organism in pure culture. Fourth, the cultured organism must cause disease when introduced into a healthy individual and must be able to be re-isolated from the new host.

Subsequently, it became clear that this is an oversimplified view of host–pathogen interactions, in that most pathogens cause disease across a spectrum, from subclinical infection to severe disease, depending on host factors (for example, the function of the immune system) and bacterial factors (for example, strain-to-strain variation in colonization and virulence factors). In addition, some pathogens cannot be grown in the laboratory, and some cause disease only in partnership with other organisms.

A molecular version of Koch's postulates has been devised by Stanley Falkow, in an attempt to provide a definition of the term 'virulence factor'[98]. This new version has three criteria. First, the potential virulence factor should be found in all pathogenic strains of a species but be absent from their non-pathogenic relatives. Second, specific inactivation of the relevant gene(s) should attenuate virulence in an appropriate animal model. Third, subsequent reintroduction of the gene should restore virulence in the animal model.

Similar to the original Koch's postulates, however, there are problems if these 'molecular Koch's postulates' are applied uncritically. These new postulates rest on the assumption that there is an essential distinction between pathogens and non-pathogens, but bacteria often have different roles in different circumstances. For example, uropathogenic *Escherichia coli* function as commensal microorganisms in the human gut but as pathogens in the human bladder, and enterohaemorrhagic *E. coli* function as commensal microorganisms in the bovine gut but are pathogens in the human gut. Similarly, *Yersinia pestis* is a pathogen of mice and fleas, but the virulence factors are likely to differ in each host.

A key contribution of genomics to this debate is to highlight the tension between the first of Falkow's postulates (virulence factors defined by using comparative genomics) and the rest of the postulates (virulence factors defined by using genetic techniques and models of infection). If the first postulate is enforced — that is, any factors that are common to pathogens and non-pathogens cannot be virulence factors — then some pathogens do not have any virulence factors. If the first postulate is ignored, then many 'virulence factors' turn up in non-pathogens (Table 2).

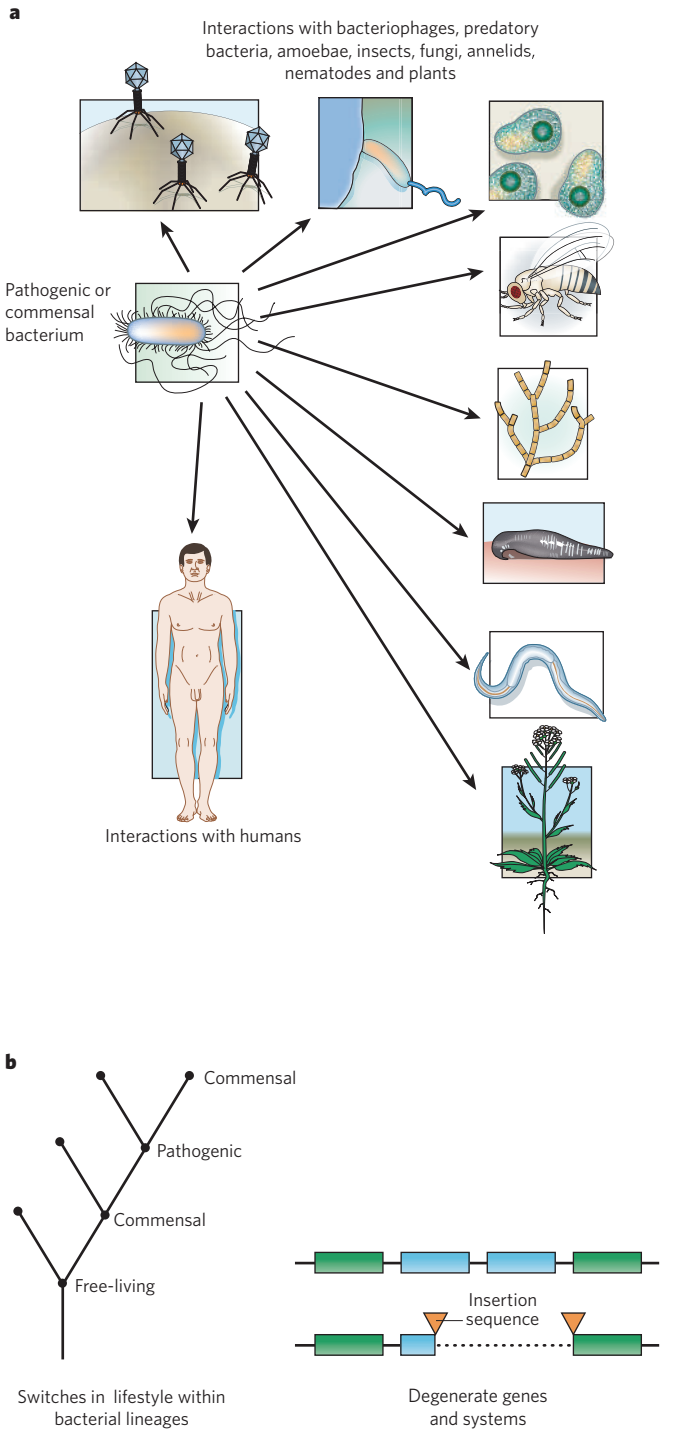


Figure 2 | The eco-evo view of bacterial pathogenomics. a, Pathogenic bacteria and commensal bacteria often share their habitats with bacteriophages, other bacteria, amoebae, insects, nematodes, annelids (such as leeches), fungi, plants and mammals (such as humans). This mixed ecology is a considerable driving force in the evolution of these microorganisms. In this context, it is not surprising that genes encoding 'virulence factors' are found in both human pathogens and non-pathogens. **b**, In addition, consideration of the evolutionary history of a pathogen might be needed to explain some of the features of its genome. Within bacterial genomes, it is common to find remnants of genes or gene clusters that presumably provided an adaptive advantage in the past but are now non-functional (indicated in blue). Also, it should be considered that a microorganism that is pathogenic now might at one time have been a commensal microorganism, and vice versa (indicated by the phylogenetic tree).

genome decay is that of human mitochondria, which belong to the α-proteobacterial lineage and retain a tiny, 17-kilobase, genome (arguably the first bacterial genome to be sequenced[47]).

Less common than genome decay, but more marked in its consequences, is positive selection for gene loss. This occurs as a newly emerged pathogen adapts to its niche and forms part of a process known as pathoadaptation. Pathoadaptation can involve any changes that refine newly formed virulence mechanisms. One example is glucosylation of the surface molecule lipopolysaccharide, which optimizes the exposure of the type III secretion apparatus of *Shigella flexneri*[48]. Pathoadaptation also encompasses gene loss, although it might seem counter-intuitive that losing genes can specifically improve the fitness of a bacterium *in vivo* and make it more pathogenic. The best-known example occurs among the shigellae: loss of the gene *cadA* (which encodes the enzyme lysine decarboxylase) provides a selective advantage in the intracellular niche, because the product of lysine-decarboxylase activity, cadaverine, inhibits the plasmid-encoded virulence factors of these bacteria[49]. Genome sequencing has shown that the genetic mechanisms underlying loss of *cadA* vary between *Shigella* lineages, thus providing an example of convergent evolution in bacterial genomes.

Intriguingly, several recently emerged pathogens (including *Bordetella pertussis*, *B. mallei*, *Y. pestis*, all *Shigella* lineages and some *E. coli* O157 lineages) have independently lost flagellar motility during the transition to a new virulence-associated lifestyle. This suggests that these bacteria are subject to a common pathoadaptive selective pressure, but it is unclear whether the driving force is loss of a target (the protein flagellin) recognized by both the innate immune response and the adaptive immune response in mammals or changes in bacterial metabolism that occur concurrently[50].

An 'eco-evo' perspective on host-pathogen interactions

A glance at the post-genomic landscape shows that our previous knowledge of the ecology and evolution of bacterial pathogenesis was limited. New findings mean that previous assumptions need to be questioned and terms need to be redefined. Among genetically variable bacterial species, it is now clear that a single strain rarely typifies an entire species, particularly because genomics has provided compelling evidence that commonly used laboratory strains (for example, *E. coli* K-12, *S. enterica* serovar Typhimurium LT2, *Pseudomonas aeruginosa* PAO1 and *S. aureus* COL) have undergone marked genotypic and phenotypic changes during their descent from the ancestral free-living isolate[51,52].

Similarly, the readily available bacterial genome-sequence data have challenged the simplistic views that a bacterial pathogen can be understood solely by identifying its virulence factors and that pathogens always evolve from non-pathogens by acquiring virulence genes on plasmids, bacteriophages or pathogenicity islands. Instead, genomics has helped to blur the distinction between pathogens and non-pathogens and between virulence factors and colonization factors. And it has catalysed a copernican shift in how host–pathogen interactions are viewed, a shift away from an anthropocentric focus towards a broader perspective that places interactions between eukaryophilic bacteria and eukaryotes in a wider ecological and evolutionary context (Fig. 2). Inherent in this 'eco–evo' perspective is the need to identify the selective advantages of virulence factors in the broader lifestyle of the pathogen. In addition, 'evolutionary narratives' that interweave genomic changes with ecological shifts can now be constructed. For example, genomic comparisons allow a reconstruction of how the plague bacillus, *Y. pestis* (a rodent and flea pathogen that is occasionally transmitted to humans), evolved from a gastrointestinal pathogen (*Yersinia pseudotuberculosis*) in an evolutionary blink of an eye (about 10,000 years), through the processes of gene gain, loss and rearrangement[8,53,54].

A more fundamental consequence of the eco–evo view is that it is now expected that what, at first, seem to be virulence factors are encoded in the genomes of 'non-pathogens' (Table 2). There are several reasons for this. First, it is now clear, both from genomic and pathogenesis studies, that pathogens, commensal microorganisms and symbionts rely

Table 2 | Examples of 'virulence systems' encoded in the genomes of both pathogenic and non-pathogenic bacteria

Virulence factor	Role in virulence	Homologues in non-pathogens	Potential explanation for presence in non-pathogens	References
Type III secretion system	Role in infection with many human pathogens, including chlamydiae, salmonellae, shigellae and yersiniae	Remnants of type III secretion systems and effectors in commensal strains of *Escherichia coli*, including the laboratory strain *E. coli* K-12	Had a role in a former niche (a degenerate system)	29, 36
		Type III secretion systems in environmental bacteria: for example, *Myxococcus xanthus*, *Verrucomicrobium spinosum*, *Desulfovibrio vulgaris* and non-pathogenic *Yersinia* spp.	Mediate uncharacterized interactions with nematodes, and amoebae and other microscopic eukaryotes in terrestrial and aquatic environments	67
		Type III secretion systems in symbiotic bacteria: for example, *Photorhabdus luminescens*, *'Hamiltonella defensa'*, *Aeromonas veronii*, *Sodalis glossinidius* and *Protochlamydia amoebophila*	Mediate symbiosis with plants, nematodes, leeches, insects and amoebae	67–73
Type VI secretion system	Role in infection with *Vibrio cholerae* or *Pseudomonas aeruginosa*	Type VI secretion systems in environmental bacteria: for example, *Rhodopirellula baltica*, *Hahella chejuensis* and *Oceanobacter* sp.	Mediate uncharacterized interactions with nematodes, and amoebae and other microscopic eukaryotes in aquatic environments	74, 75
ESAT6 and associated Esx secretion system	Key virulence determinant of *Mycobacterium tuberculosis* and *Staphylococcus aureus*, and major attenuating factor in the bacillus Calmette–Guérin (BCG) vaccine against tuberculosis	Esx gene clusters in *Bacillus subtilis*, *Bacillus licheniformis*, *Bacillus halodurans*, *Clostridium acetobutylicum*, *Listeria innocua* and *Streptomyces coelicolor*	Mediate uncharacterized interactions with nematodes, and amoebae and other microscopic eukaryotes in terrestrial and aquatic environments, or involved in conjugative transfer of plasmids	76–81
Specific invasion genes (for example, *yijP*, *ibeB* and *ompA*)	Contribute to invasion of *E. coli* in animal models of meningitis	Invasion genes in commensal strains of *E. coli*, including the laboratory strain *E. coli* K-12	Are a short-sighted local adaptation that does not contribute to transmission (a dead-end trait)	82–84

on similar strategies and molecular systems in their interactions with eukaryotic hosts (for example, phase variation)[21,55]. Second, it is also understood that bacteria sometimes produce virulence factors that provided an advantage only in a previous, now non-existent, niche. Last, it has also become evident that many bacterial pathogens infect humans only incidentally and often produce virulence factors that are active against non-mammalian adversaries as diverse as plants, insects, protozoans, nematodes, predatory bacteria and bacteriophages. Inherent in this view is the realization that many bacterial virulence factors have been shaped by evolutionary forces outside the context of human–pathogen interactions, and only by studying these forces can the emergence of human infections be understood.

Enterohaemorrhagic *E. coli* strains, particularly *E. coli* O157, provide a compelling test case for the eco–evo view. *E. coli* O157 is a rare but devastating pathogen of humans, but it is also a common commensal microorganism of the bovine gut. Genomic comparisons have helped to explain how this pathogen has evolved from a non-pathogenic ancestor by acquiring virulence factors encoded on various mobile elements (for example, Shiga toxin, which is encoded on a bacteriophage, and the type III secretion system encoded on the LEE pathogenicity island)[56]. However, recent studies have shown that a pilus-adherence factor that is crucial to the virulence of *E. coli* O157 in humans is also carried by commensal strains of *E. coli*[57]. Also, similarly to many commensal strains of *E. coli*, *E. coli* O157 carries remnants of a gene cluster (ETT2) encoding a virulence-associated secretion system that is now thought to be inactive[36]. It is only through an eco–evo view that the evolution and transmission of *E. coli* O157, and why it produces such lethal virulence factors, might be understood. One potential explanation is that the 'virulence factors' of *E. coli* O157, such as Shiga toxin, help it to colonize the bovine gut. However, the evidence for this hypothesis is equivocal at best[58]. Instead, a recent study shows that the Shiga-toxin-encoding bacteriophage increases bacterial survival in the presence of a grazing ciliate, *Tetrahymena pyriformis*, indicating that interactions with non-mammalian adversaries might have driven the evolution of this virulence factor[59]. Similar points can be made about many other pathogens: for example, it has long been known that the pathogenesis of legionellosis in humans relies on mechanisms that legionellae use to subvert amoebae[60]. Clearly, in the post-genomic era, even for important human pathogens, humans can no longer be considered to be the centre of the bacterial universe.

Future challenges

The clearest challenge for future studies of bacterial pathogenomics is coping with the flood of new data unleashed by the arrival of affordable and quick, 'next-generation', sequencing technologies[61]. Now that the cost of sequencing bacterial genomes fits comfortably within the budget of a standard research project grant, it is set to become an integral and routine part of research programmes. Therefore, within the next decade, tens of thousands of bacterial genomes will be sequenced[62]. And the focus will shift from the mechanics of generating sequence data to the problems of analysing it, creating an urgent need for better ways to compare and visualize genomic data. Also, there are likely to be many more incompletely sequenced genomes, as the efficiency of the finishing stage of a genome project lags behind the rapid pace of next-generation whole-genome shotgun sequencing.

For genetically uniform species, particularly those with potential as bioterrorism agents, the resequencing of hundreds of isolates (for example, by using tiling arrays) will drive forward forensic genomics. For more variable species, such as *E. coli*, *S. enterica* and *S. aureus*, the focus will be on defining the extent of the pan-genome and on developing improved approaches to understanding epidemiology, particularly for those that cause hospital-acquired infections. The most important challenge will still be to add functional relevance to genome sequences, a challenge that will continue to drive the application of high-throughput 'omics' approaches to the study of virulence.

Furthermore, sequencing the genomes of environmental organisms and carrying out metagenomic surveys of diverse environments will provide not only an improved understanding of microbial biodiversity but also insight into the evolution of bacterial factors that are involved in human disease[63,64]. Metagenomic surveys of eukaryote-associated bacterial communities will strengthen our understanding of the ecology of bacterial infections (for example, the micro-ecological changes that accompany antibiotic-associated diarrhoea) and help to shed light on the pathogenesis of polymicrobial infections, such as those that cause periodontal disease and bacterial vaginosis[65]. Similarly, studying the metagenomics of bacteriophage populations will help to unravel the connections between these mobile elements and the evolution of virulence.

In addition to the genomic technologies discussed here, evolutionary game theory will need to be applied so that the complex interactions between bacteriophages, the virulence factors that they encode, the bacteria that they infect and the eukaryotic targets of their virulence

Glossary

Bacteriophage A virus that infects bacteria. A bacteriophage can either lyse a cell or integrate into its genome.

Commensal microorganism A microorganism that benefits from living in close contact with a human or animal but has no direct beneficial or detrimental effects on its host.

Core genome The set of genes found in all members of a single species.

Eco–evo perspective A perspective in which organisms are evaluated broadly in the light of evolution and ecology, rather than narrowly by the constraints of their behaviour in the laboratory or in human infection.

Eukaryophilic A term applied to any bacterium that interacts with eukaryotes in its natural environmental niche. It does not specify whether the interaction is pathogenic or symbiotic.

Finishing stage The final phase in a genome-sequencing project, in which all gaps between contigs are closed and all ambiguities are resolved. This is much more labour-intensive than the shotgun phase, so the sequencing of many genomes is now left unfinished.

GC skew A measure that reflects bias for guanine bases on the leading strand of DNA and cytosine bases on the lagging strand.

Genome The complete set of genetic information in an organism. In bacteria, this includes the chromosome(s) and plasmids.

Horizontal gene transfer Any process in which an organism transfers genetic material to another cell that is not its offspring. This process is in contrast to vertical gene transfer, which is much more common and occurs when genetic material is passed from parent to offspring or, more generally, from ancestor to descendent.

Insertion sequence The simplest type of transposable element in bacteria. It contains only the genes required for its own transposition.

Metagenomics The high-throughput study of sequences from multiple genomes recovered from environmental samples that contain mixed populations.

Next-generation sequencing A set of novel approaches to DNA sequencing that dispenses with the need to create libraries of cloned sequences in bacteria and holds the promise of providing faster and cheaper sequencing.

Pan-genome The set of all genes found in members of a single species.

Pathoadaptation The genetic changes that occur after transition to a new pathogenic lifestyle and ensure that the bacterium becomes fitter in its new host niche. These can include changes in the sequences of genes, alterations in gene expression and loss of genes.

Pathogen An organism that can cause disease in another organism.

Pathogenicity island A cluster of genes acquired by horizontal gene transfer that encodes products that contribute to virulence.

Phase variation A spontaneous genetically defined switch between expression of alternative, usually surface-associated, proteins that occurs at a high frequency.

Plasmid An extrachromosomal DNA molecule that can replicate autonomously within a bacterial cell.

Resequencing The application of genome sequencing to a close relative of a strain that has already been sequenced. The availability of a template genome greatly facilitates the finishing stage of sequencing. In contrast to resequencing, *de novo* sequencing, in which a large proportion of the genome is novel sequence, is much more challenging.

Single-nucleotide polymorphism A variation between two genomes that involves a single base pair.

Symbiont An organism that lives in close contact with another organism (for example, a bacterium and a eukaryote) in a relationship in which both partners benefit. Endosymbionts are symbionts that live within the body or cells of another organism.

Type III secretion One of several processes by which bacteria export proteins to the external environment. Type III secretion is required for the biosynthesis of the main organelle of motility in bacteria, the flagellum. It is also responsible for the translocation of bacterial effector proteins from pathogenic or symbiotic bacteria to the cytoplasm of their eukaryotic partners, where these proteins subvert eukaryotic-cell functions to the advantage of the bacterium.

Virulence factor A factor that is produced by a pathogen and required for it to cause disease. It should be noted that defining this term precisely is difficult.

Whole-genome shotgun sequencing An approach to determining the sequence of a genome in which the genome is broken into numerous small fragments. These fragments are then sequenced en masse. The individual sequences are assembled into larger sequences (known as contigs) that correspond to substantial portions of the genome. Typically, more than 99% of the genome can be sequenced using this approach, before the finishing stage.

factors can be understood. Similar approaches will also be needed to solve the conundrum of invasive disease: for example, to explain why meningococci cause meningitis, despite the fact that the disease has no role in the transmission of the bacteria[66]. Evolutionary systems-biology approaches will also be useful for understanding the evolution and regulation of complex virulence systems, the interactions between pathogens and their host, and the co-evolution of their genomes.

The first decade of bacterial genomics has afforded unprecedented insights into the evolution of virulence. The next decade holds the promise of being even more rewarding as the new eco–evo view of host–pathogen interactions draws on ever more genome and metagenome sequences. ■

1. Read, T. D. *et al.* Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* **296,** 2028–2033 (2002).
2. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc. Natl Acad. Sci. USA* **102,** 13950–13955 (2005).
3. Hayashi, T. *et al.* Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8,** 11–22 (2001); erratum **8,** 96 (2001).
4. Kuroda, M. *et al.* Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* **357,** 1225–1240 (2001).
5. Fraser-Liggett, C. M. Insights on biology and evolution from microbial genome sequencing. *Genome Res.* **15,** 1603–1610 (2005).
6. Lawrence, J. G. Horizontal and vertical gene transfer: the life history of pathogens. *Contrib. Microbiol.* **12,** 255–271 (2005).
7. Raskin, D. M., Seshadri, R., Pukatzki, S. U. & Mekalanos, J. J. Bacterial genomics and pathogen evolution. *Cell* **124,** 703–714 (2006).
8. Wren, B. W. The yersiniae — a model genus to study the rapid evolution of bacterial pathogens. *Nature Rev. Microbiol.* **1,** 55–64 (2003).
9. Pearson, T. *et al.* Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc. Natl Acad. Sci. USA* **101,** 13536–13541 (2004).
10. Touchman, J. W. *et al.* A North American *Yersinia pestis* draft genome aequence: SNPs and phylogenetic analysis. *PLoS ONE* **2,** e220 (2007).
11. Gutacker, M. M. *et al.* Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* **162,** 1533–1543 (2002).
12. Monot, M. *et al.* On the origin of leprosy. *Science* **308,** 1040–1042 (2005).

13. Hayashi, K. *et al.* Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Syst. Biol.* **2,** doi:10.1038/msb4100049 (2006).
14. Zhang, W. *et al.* Probing genomic diversity and evolution of *Escherichia coli* O157 by single nucleotide polymorphisms. *Genome Res.* **16,** 757–767 (2006).
15. Chaudhuri, R. R. *et al.* Genome sequencing shows that European isolates of *Francisella tularensis* subspecies *tularensis* are almost identical to US laboratory strain Schu S4. *PLoS ONE* **2,** e352 (2007).
16. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).
17. Romero, C. M. *et al.* Genome sequence alterations detected upon passage of *Burkholderia mallei* ATCC 23344 in culture and in mammalian hosts. *BMC Genomics* **7,** 228 (2006).
18. Moxon, R., Bayliss, C. & Hood, D. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* **40,** 307–333 (2006).
19. van der Woude, M. W. & Baumler, A. J. Phase and antigenic variation in bacteria. *Clin. Microbiol. Rev.* **17,** 581–611 (2004).
20. Parkhill, J. *et al.* The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403,** 665–668 (2000).
21. Cerdeno-Tarraga, A. M. *et al.* Extensive DNA inversions in the *B. fragilis* genome control variable gene expression. *Science* **307,** 1463–1465 (2005).
22. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15,** 589–594 (2005).
23. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393,** 537–544 (1998).
24. Bruggemann, H., Cazalet, C. & Buchrieser, C. Adaptation of *Legionella pneumophila* to the host environment: role of protein secretion, effectors and eukaryotic-like proteins. *Curr. Opin. Microbiol.* **9,** 86–94 (2006).
25. Salanoubat, M. *et al.* Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415,** 497–502 (2002).
26. Turner, S. M. *et al.* Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. *J. Clin. Microbiol.* **44,** 4528–4536 (2006).
27. Freeman, V. J. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J. Bacteriol.* **61,** 675–688 (1951).
28. Brussow, H., Canchaya, C. & Hardt, W. D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68,** 560–602 (2004).
29. Tobe, T. *et al.* An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc. Natl Acad. Sci. USA* **103,** 14941–14946 (2006).
30. Ohnishi, M., Kurokawa, K. & Hayashi, T. Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol.* **9,** 481–485 (2001).
31. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8,** 504–508 (2000).
32. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nature Rev. Microbiol.* **2,** 414–424 (2004).
33. Zhang, L. *et al.* Regulators encoded in the *Escherichia coli* type III secretion system 2 gene cluster influence expression of genes within the locus for enterocyte effacement in enterohemorrhagic *E. coli* O157:H7. *Infect. Immun.* **72,** 7282–7293 (2004).
34. Nakanishi, N. *et al.* ppGpp with *DksA* controls gene expression in the locus of enterocyte effacement (LEE) pathogenicity island of enterohaemorrhagic *Escherichia coli* through activation of two virulence regulatory genes. *Mol. Microbiol.* **61,** 194–205 (2006).
35. Laaberki, M. H., Janabi, N., Oswald, E. & Repoila, F. Concert of regulators to switch on LEE expression in enterohemorrhagic *Escherichia coli* O157:H7: interplay between Ler, GrlA, HNS and RpoS. *Int. J. Med. Microbiol.* **296,** 197–210 (2006).
36. Ren, C. P. *et al.* The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. *J. Bacteriol.* **186,** 3547–3560 (2004).
37. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17,** 589–596 (2001).
38. Ren, C. P., Beatson, S. A., Parkhill, J. & Pallen, M. J. The *Flag-2* locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*. *J. Bacteriol.* **187,** 1430–1440 (2005).
39. Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409,** 1007–1011 (2001).
40. Parkhill, J. *et al.* Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413,** 523–527 (2001).
41. Parkhill, J. *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413,** 848–852 (2001).
42. Siguier, P., Filee, J. & Chandler, M. Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.* **9,** 526–531 (2006).
43. Wernegreen, J. J. For better or worse: genomic consequences of intracellular mutualism and parasitism. *Curr. Opin. Genet. Dev.* **15,** 572–583 (2005).
44. Andersson, S. G. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396,** 133–140 (1998).
45. Perez-Brocal, V. *et al.* A small microbial genome: the end of a long symbiotic relationship? *Science* **314,** 312–313 (2006).
46. Nakabachi, A. *et al.* The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314,** 267 (2006).
47. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).
48. West, N. P. *et al.* Optimization of virulence functions through glucosylation of *Shigella* LPS. *Science* **307,** 1313–1317 (2005).
49. Maurelli, A. T. Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. *FEMS Microbiol. Lett.* **267,** 1–8 (2007).
50. Leatham, M. P. *et al.* Mouse intestine selects nonmotile *flhDC* mutants of *Escherichia coli* MG1655 with increased colonizing ability and better utilization of carbon sources. *Infect. Immun.* **73,** 8039–8049 (2005).
51. Fux, C. A., Shirtliff, M., Stoodley, P. & Costerton, J. W. Can laboratory reference strains mirror 'real-world' pathogenesis? *Trends Microbiol.* **13,** 58–63 (2005).
52. Hobman, J. L., Penn, C. W. & Pallen, M. J. Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Mol. Microbiol.* **64,** 881–885 (2007).
53. Achtman, M. *et al.* Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proc. Natl Acad. Sci. USA* **101,** 17837–17842 (2004).
54. Achtman, M. *et al.* *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA* **96,** 14043–14048 (1999).
55. van der Woude, M. W. Re-examining the role and random nature of phase variation. *FEMS Microbiol. Lett.* **254,** 190–197 (2006).
56. Wick, L. M., Qi, W., Lacher, D. W. & Whittam, T. S. Evolution of genomic content in the stepwise emergence of *Escherichia coli* O157:H7. *J. Bacteriol.* **187,** 1783–1791 (2005).
57. Rendón, M. A. *et al.* Commensal and pathogenic *Escherichia coli* use a common pilus adherence factor for epithelial cell colonization. *Proc. Natl Acad. Sci. USA* **104,** 10637–10642 (2007).
58. Sheng, H., Lim, J. Y., Knecht, H. J., Li, J. & Hovde, C. J. Role of *Escherichia coli* O157:H7 virulence factors in colonization at the bovine terminal rectal mucosa. *Infect. Immun.* **74,** 4685–4693 (2006).
59. Meltz Steinberg, K. & Levin, B. R. Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 Shiga toxin-encoding prophage. *Proc. R. Soc. B* **274,** 1921–1929 (2007).
60. Albert-Weissenberger, C., Cazalet, C. & Buchrieser, C. *Legionella pneumophila* — a human pathogen that co-evolved with fresh water protozoa. *Cell. Mol. Life Sci.* **64,** 432–448 (2007).
61. Hall, N. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* **210,** 1518–1525 (2007).
62. Field, D., Wilson, G. & van der Gast, C. How do we compare hundreds of bacterial genomes? *Curr. Opin. Microbiol.* **9,** 499–504 (2006).
63. Rusch, D. B. *et al.* The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* **5,** e77 (2007).
64. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
65. Brogden, K. A., Guthmiller, J. M. & Taylor, C. E. Human polymicrobial infections. *Lancet* **365,** 253–255 (2005).
66. Meyers, L. A., Levin, B. R., Richardson, A. R. & Stojiljkovic, I. Epidemiology, hypermutation, within-host evolution and the virulence of *Neisseria meningitidis*. *Proc. R. Soc. B* **270,** 1667–1677 (2003).
67. Pallen, M. J., Beatson, S. A. & Bailey, C. M. Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. *FEMS Microbiol. Rev.* **29,** 201–229 (2005).
68. Ffrench-Constant, R. H. *et al.* A genomic sample sequence of the entomopathogenic bacterium *Photorhabdus luminescens* W14: potential implications for virulence. *Appl. Environ. Microbiol.* **66,** 3310–3329 (2000).
69. Moran, N. A., Degnan, P. H., Santos, S. R., Dunbar, H. E. & Ochman, H. The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proc. Natl Acad. Sci. USA* **102,** 16919–16926 (2005).
70. Silver, A. C. *et al.* Interaction between innate immune cells and a bacterial type III secretion system in mutualistic and pathogenic associations. *Proc. Natl Acad. Sci. USA* **104,** 9481–9486 (2007).
71. Skorpil, P. *et al.* NopP, a phosphorylated effector of *Rhizobium* sp. strain NGR234, is a major determinant of nodulation of the tropical legumes *Flemingia congesta* and *Tephrosia vogelii*. *Mol. Microbiol.* **57,** 1304–1317 (2005).
72. Horn, M. *et al.* Illuminating the evolutionary history of chlamydiae. *Science* **304,** 728–730 (2004).
73. Dale, C., Young, S. A., Haydon, D. T. & Welburn, S. C. The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. *Proc. Natl Acad. Sci. USA* **98,** 1883–1888 (2001).
74. Pukatzki, S. *et al.* Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proc. Natl Acad. Sci. USA* **103,** 1528–1533 (2006).
75. Mougous, J. D. *et al.* A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. *Science* **312,** 1526–1530 (2006).
76. Pym, A. S., Brodin, P., Brosch, R., Huerre, M. & Cole, S. T. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. *Mol. Microbiol.* **46,** 709–717 (2002).
77. Lewis, K. N. *et al.* Deletion of RD1 from *Mycobacterium tuberculosis* mimics bacille Calmette–Guerin attenuation. *J. Infect. Dis.* **187,** 117–123 (2003).
78. Brodin, P. *et al.* Dissection of ESAT-6 system 1 of *Mycobacterium tuberculosis* and impact on immunogenicity and virulence. *Infect. Immun.* **74,** 88–98 (2006).
79. Pallen, M. J. The ESAT-6/WXG100 superfamily — and a new Gram-positive secretion system? *Trends Microbiol.* **10,** 209–212 (2002).
80. Desvaux, M. *et al.* Genomic analysis of the protein secretion systems in *Clostridium acetobutylicum* ATCC 824. *Biochim. Biophys. Acta* **1745,** 223–253 (2005).
81. Burts, M. L., Williams, W. A., DeBord, K. & Missiakas, D. M. EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of *Staphylococcus aureus* infections. *Proc. Natl Acad. Sci. USA* **102,** 1169–1174 (2005).
82. Huang, S. H. *et al.* Identification and characterization of an *Escherichia coli* invasion gene locus, *ibeB*, required for penetration of brain microvascular endothelial cells. *Infect. Immun.* **67,** 2103–2109 (1999).
83. Huang, S. H., Stins, M. F. & Kim, K. S. Bacterial penetration across the blood–brain barrier during the development of neonatal meningitis. *Microbes Infect.* **2,** 1237–1244 (2000).
84. Huang, S. H., Wan, Z. S., Chen, Y. H., Jong, A. Y. & Kim, K. S. Further characterization of *Escherichia coli* brain microvascular endothelial cell invasion gene *ibeA* by deletion, complementation, and protein expression. *J. Infect. Dis.* **183,** 1071–1078 (2001).
85. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
86. Parkhill, J. The importance of complete genome sequences. *Trends Microbiol.* **10,** 219–220 (2002).
87. Dorer, M. S. & Isberg, R. R. Non-vertebrate hosts in the analysis of host–pathogen interactions. *Microbes Infect.* **8,** 1637–1646 (2006).

88. Hilbi, H., Weber, S. S., Ragaz, C., Nyfeler, Y. & Urwyler, S. Environmental predators as models for bacterial pathogenesis. *Environ. Microbiol.* **9,** 563–575 (2007).
89. Maiden, M. C. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60,** 561–588 (2006).
90. Lindstedt, B. A. Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. *Electrophoresis* **26,** 2567–2582 (2005).
91. Dorrell, N., Hinchliffe, S. J. & Wren, B. W. Comparative phylogenomics of pathogenic bacteria by microarray analysis. *Curr. Opin. Microbiol.* **8,** 620–626 (2005).
92. Champion, O. L. *et al.* Comparative phylogenomics of the food-borne pathogen *Campylobacter jejuni* reveals genetic markers predictive of infection source. *Proc. Natl Acad. Sci. USA* **102,** 16043–16048 (2005).
93. Renesto, P. *et al.* Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. *Lancet* **362,** 447–449 (2003).
94. Wacker, M. *et al.* N-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. *Science* **298,** 1790–1793 (2002).
95. Mora, M., Donati, C., Medini, D., Covacci, A. & Rappuoli, R. Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. *Curr. Opin. Microbiol.* **9,** 532–536 (2006).
96. Jenner, R. G. & Young, R. A. Insights into host responses against pathogens from transcriptional profiling. *Nature Rev. Microbiol.* **3,** 281–294 (2005).
97. Hill, A. V. Aspects of genetic susceptibility to human infectious diseases. *Annu. Rev. Genet.* **40,** 469–486 (2006).
98. Falkow, S. Molecular Koch's postulates applied to microbial pathogenicity. *Rev. Infect. Dis.* **10** (suppl. 2), S274–S276 (1988).

Acknowledgements We thank L. Snyder, J. Kelly, D. Baker, L. Bingle and S. Andersson for critical reading of the manuscript. We acknowledge the Biotechnology and Biological Sciences Research Council for funding numerous genomic research projects in our laboratories, and the Wellcome Trust (particularly the Wellcome Trust Sanger Institute) for facilitating bacterial genome sequencing in the United Kingdom. This article is dedicated to the memory of C. A. Hart.

Author Information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to the authors (m.pallen@bham.ac.uk; brendan.wren@lshtm.ac.uk).

A very faint core-collapse supernova in M85

Arising from: S. R. Kulkarni *et al. Nature* **447**, 458–460 (2007)

An anomalous transient in the early Hubble-type (S0) galaxy Messier 85 (M85) in the Virgo cluster was discovered by Kulkarni *et al.*[1] on 7 January 2006 that had very low luminosity (peak absolute *R*-band magnitude M_R of about −12) that was constant over more than 80 days, red colour and narrow spectral lines, which seem inconsistent with those observed in any known class of transient events. Kulkarni *et al.*[1] suggest an exotic stellar merger as the possible origin. An alternative explanation is that the transient in M85 was a type II-plateau supernova of extremely low luminosity, exploding in a lenticular galaxy with residual star-forming activity. This intriguing transient might be the faintest supernova that has ever been discovered.

Kulkarni *et al.*[1] suggest that this event (labelled M85 OT2006-1) was an anomalous luminous red nova, possibly generated in a stellar merger. An overall similarity with other transients (M31 RV, V4332 Sgr and V838 And) was also proposed[2], in spite of the fact that M85 OT2006-1 is significantly brighter (by two to three magnitudes). We propose an alternative scenario in which M85 OT2006-1 is an extreme case of a low-luminosity type II-plateau supernova[3,4], about 1.5 magnitudes fainter than the faintest low-luminosity type II-plateau supernova discovered so far (SN1999br)[4,5].

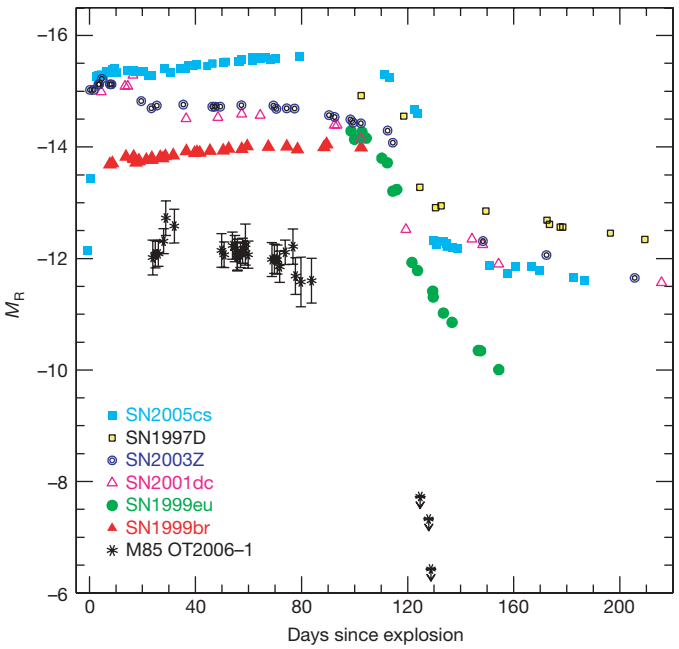


Figure 1 | Absolute *R*-band magnitude light curves for M85 OT2006-1 and low-luminosity type II-plateau supernovae. The data for the supernova sample, adopted explosion epochs, interstellar extinction and distance moduli (calibrated with $H_0 = 72\ \mathrm{km\ s^{-1}\ mpc^{-1}}$) are from refs 4 and 6. Unpublished data of the faint type II-plateau supernova SN2003Z, with distance modulus $(m - M) = 32.64$ (where m is the apparent magnitude) and interstellar extinction $E(B-V) = 0.04$, are also displayed. For M85 OT2006-1, we make use of a direct estimate of the surface brightness fluctuation distance to M85 (ref. 15), giving $(m - M) = 31.33 \pm 0.14$ (18.5 ± 1.2 mpc), adopting the same interstellar absorption estimate ($A_R = 0.4$) as in Kulkarni *et al.*[1]. With an average plateau magnitude $R = 19.65$, we obtain an absolute magnitude of $M_R = -12.1$. Low-luminosity type II-plateau supernovae cover an extended range of luminosity, and thus the light curve of M85 OT2006-1 (considering the photometric errors from Kulkarni *et al.*[1]) is consistent with that of an extremely faint type II-plateau supernova.

The morphology of its light curve[1] is very similar to that of an low-luminosity type II-plateau supernova , in terms of the plateau duration (typically 80–110 days) and the low peak luminosity ($-15 < M_V < -13$; refs 4–6 and Fig. 1). The spectra of M85 OT2006-1 show a red continuum and very narrow lines (H I, Ca II, Ba II, Fe II and possibly other species; Fig. 2), which are commonly observed in low-luminosity type II-plateau supernovae during the hydrogen envelope recombination phase[3–6]. Some features show P Cygni profiles, others a double component in emission. In particular, Hα shows a very narrow line ($v \approx 350\ \mathrm{km\ s^{-1}}$; ref. 1) on the top of a broader ($v \approx 800\ \mathrm{km\ s^{-1}}$) component (Fig. 2b), which is consistent with the typical expansion velocities of known low-luminosity type II-plateau supernovae[4]. The properties of this subclass of supernovae are explained by invoking low explosion energies ($\ll 10^{51}$ erg) and very small amounts of ejected radioactive ^{56}Ni ($\ll 10^{-2} M_\odot$, where $M_\odot$ is the mass of the Sun; ref. 4).

Core-collapse supernovae are normally observed in spirals, whereas M85 is a lenticular galaxy (S0). However, traces of recent episodes of star formation are observed in many local early-type galaxies, in the form of faint nebular emission lines, tidal tails, dust lanes, H I gas, blue colour and radio emission. Hence the fact that M85 OT2006-1 occurred in an S0 galaxy does not categorically exclude a relatively massive progenitor star. It has been suggested that moderate-mass stars ($8–12 M_\odot$) can explode as low-luminosity type II-plateau supernovae[7–10], and it is certainly possible that Hubble type E/S0 galaxies host stellar populations of such masses. Indeed, about a dozen core-collapse supernovae have been discovered in S0 galaxies in the past five years[11]. This might well be the case for M85, which has a non-negligible radio flux ($L \approx 10^{27}\ \mathrm{erg\ s^{-1}\ Hz^{-1}}$), placing this galaxy in the faint tail of the radio galaxies[12]. Moreover, the strong nuclear Hα and the presence of an inner blue ring in M85 are indicative that recent (minor) episodes of star formation occurred[13].

One of the arguments advanced by Kulkarni *et al.*[1] against a massive star origin for M85 OT2006-1 was that no star was visible in deep pre-explosion Hubble Space Telescope images. However, the detection limit of $m_{F850LP} = 24.7$ (as quoted by Kulkarni *et al.*[1]) is not sufficient to exclude a moderately massive precursor. If it was a spectroscopic type K to M supergiant, after adopting the assumptions on the distance and the reddening by Kulkarni *et al.*[1], a luminosity limit of $\log L/L_\odot \approx 4.8$ is obtained. Using current stellar evolutionary tracks for an isolated red supergiant, we obtain an upper mass limit of $15 M_\odot$. The deep detection limit in the blue filter ($m_{F475W} = 26.8$) can be used to restrict blue supergiants (B-type, as in the case of SN1987A) to less than about $12 M_\odot$. Small star clusters of initial mass about $300 M_\odot$ (obtained using the stellar population models of Starburst99; ref. 14) would be undetected with these magnitude limits. Assuming a typical initial mass function in such a cluster, one could quite reasonably expect a few stars above the limit of $8 M_\odot$. In light of all of this, we suggest a plausible physical interpretation in which the M85 transient is an extremely faint core-collapse supernova, characterized by a very low explosion energy ($5–10 \times 10^{49}$ erg), a small amount of ^{56}Ni ($\ll 10^{-3} M_\odot$) and a total ejected mass of $6–9 M_\odot$.

A. Pastorello[1], M. Della Valle[2,3,4], S. J. Smartt[1], L. Zampieri[5], S. Benetti[4,5], E. Cappellaro[4,5], P. A. Mazzali[6,7], F. Patat[8], S. Spiro[1,5], M. Turatto[4,5] & S. Valenti[8,9]

[1]Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, UK.
e-mail: a.pastorello@qub.ac.uk
[2]INAF Osservatorio Astronomico di Arcetri, 50125 Firenze, Italy.

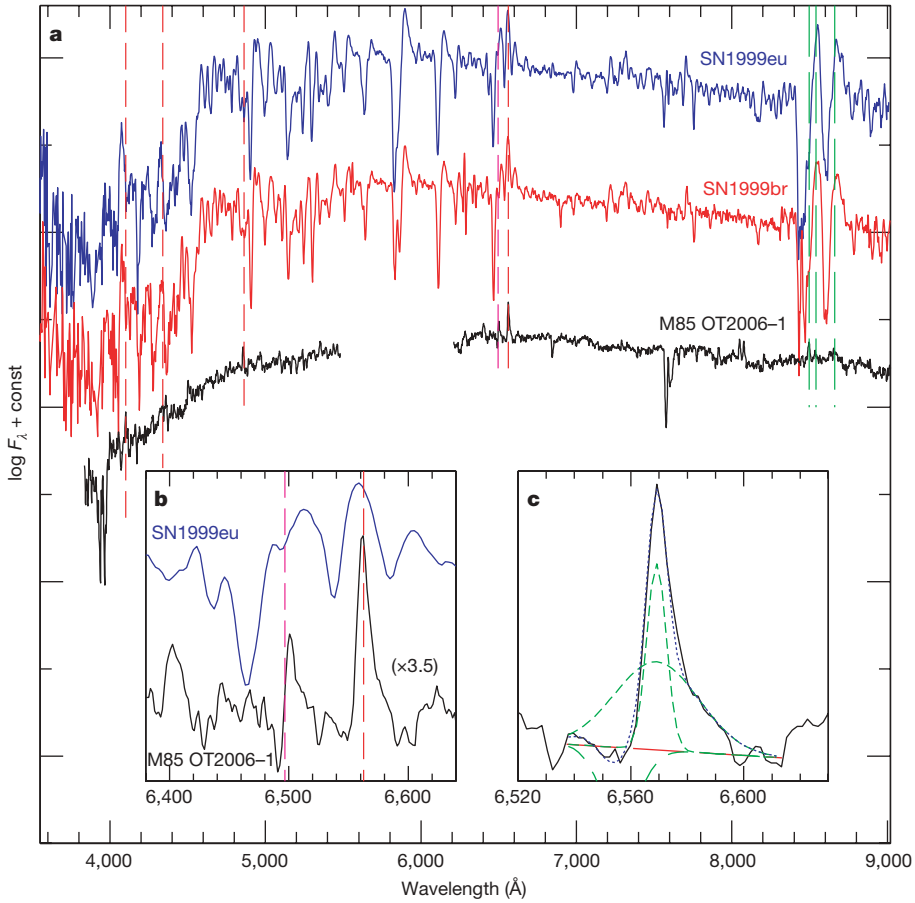


Figure 2 | Spectral properties of M85 OT2006-1 and low-luminosity type II-plateau supernovae. a, The spectrum of M85 OT2006-1 obtained on 24 February 2006 (from ref. 1, kindly provided by S. R. Kulkarni and A. Rau, de-reddened by $A_R = 0.4$ magnitudes) is compared with those of SN1999br and SN1999eu observed at the end of the plateau[4]. The position of the H Balmer lines, Ba II λ6497 and the near infrared Ca II triplet are marked by vertical dashed lines. These lines are visible in the spectrum of the transient, whereas there is no clear evidence of molecular TiO bands, characterizing spectra of red giants and the peculiar luminous red novae proposed in refs 1 and 2. F is the flux (in erg s^{-1}) scaled to an arbitrary constant. The axis units are the same for **b** and **c**. **b**, The region of the P Cygni Ba II λ6497 + Hα lines in SN1999eu and M85 OT2006-1. **c**, Three-component fit (broader P Cygni feature with $V_{exp} \approx 700$–800 km s^{-1}, plus narrow emission) to the Hα profile. The red line shows the level of the continuum, the green lines show separately the three components used to fit the line, and the dotted line is the resulting three-component fit.

[3]International Center for Relativistic Astrophysics Network, 65122 Pescara, Italy.
[4]Kavli Institute for Theoretical Physics, University of California Santa Barbara, California 93106, USA.
[5]INAF Osservatorio Astronomico di Padova, 35122 Padova, Italy.
[6]Max-Planck-Institut für Astrophysik, 85741 Garching bei München, Germany.
[7]INAF Osservatorio Astronomico di Trieste, 34131 Trieste, Italy.
[8]European Southern Observatory, 85748 Garching bei München, Germany.
[9]Dipartimento di Fisica, Università di Ferrara, 44100 Ferrara, Italy.

Received 18 July; accepted 17 August 2007.

1. Kulkarni, S. R. *et al.* An unusually brilliant transient in the galaxy Messier 85. *Nature* **447**, 458–460 (2007).
2. Rau, A., Kulkarni, S. R., Ofek, E. O. & Yan, L. Spitzer observations of the new luminous red nova M85 OT2006–1. *Astrophys. J.* **659**, 1536–1540 (2007).
3. Turatto, M. *et al.* The peculiar type II supernova 1997D: a case for a very low ^{56}Ni mass. *Astrophys. J.* **498**, L129–L133 (1998).
4. Pastorello, A. *et al.* Low-luminosity type II supernovae: spectroscopic and photometric evolution. *Mon. Not. R. Astron. Soc.* **347**, 74–94 (2004).
5. Zampieri, L. *et al.* Peculiar, low-luminosity type II supernovae: low-energy explosions in massive progenitors? *Mon. Not. R. Astron. Soc.* **338**, 711–716 (2003).
6. Pastorello, A. *et al.* SN 2005cs in M51—I. The first month of evolution of a subluminous SN II plateau. *Mon. Not. R. Astron. Soc.* **370**, 1752–1762 (2006).
7. Chugai, N. N. & Utrobin, V. P. The nature of SN 1997D: low-mass progenitor and weak explosion. *Astron. Astrophys.* **354**, 557–566 (2000).
8. Maund, J. R., Smartt, S. J. & Danziger, I. J. The progenitor of SN 2005cs in the Whirlpool galaxy. *Mon. Not. R. Astron. Soc.* **364**, L33–L37 (2005).
9. Li, W. *et al.* Identification of the red supergiant progenitor of supernova 2005cs: do the progenitors of type II-P supernovae have low mass? *Astrophys. J.* **641**, 1060–1070 (2006).
10. Kitaura, F. S., Janka, H.-T. & Hillebrandt, W. Explosions of O-Ne-Mg cores, the Crab supernova, and subluminous type II-P supernovae. *Astron. Astrophys.* **450**, 345–350 (2006).
11. The Padova-Asiago Supernova Catalogue. ⟨http://web.pd.astro.it/supern/⟩.
12. Della Valle, M. *et al.* Why are radio galaxies prolific producers of type Ia supernovae? *Astrophys. J.* **629**, 750–756 (2005).
13. Fisher, D. Kinematic profiles of SO galaxies. *Astron. J.* **113**, 950–974 (1997).
14. Leitherer, C. Starburst99: synthesis models for galaxies with active star formation. *Astrophys. J. Suppl. Ser.* **123**, 3–40 (1999).
15. Tonry, J. L. *et al.* The SBF survey of galaxy distances. IV. SBF magnitudes, colors, and distances. *Astrophys. J.* **546**, 681–693 (2001).

doi:10.1038/nature06282

HYPOTHESIS

The equilibria that allow bacterial persistence in human hosts

Martin J. Blaser[1] & Denise Kirschner[2]

We propose that microbes that have developed persistent relationships with human hosts have evolved cross-signalling mechanisms that permit homeostasis that conforms to Nash equilibria and, more specifically, to evolutionarily stable strategies. This implies that a group of highly diverse organisms has evolved within the changing contexts of variation in effective human population size and lifespan, shaping the equilibria achieved, and creating relationships resembling climax communities. We propose that such ecosystems contain nested communities in which equilibrium at one level contributes to homeostasis at another. The model can aid prediction of equilibrium states in the context of further change: widespread immunodeficiency, changing population densities, or extinctions.

When two organisms occupy the same habitat, a conflict or a series of compromises ensues. Sometimes there are elements of both, and interactions range from a 'cold-war'-type conflict to peaceful coexistence. Many of the most intense conflicts are accidental (for example, when a microbe finds itself in a niche (or host) to which it is unaccustomed), and the interactions are often short term (leading to the eradication of the microbe or the death of the host). More complex are the relationships between hosts and microbes that have evolved together, each with adaptations tied to the biology of the other, often leading to nonlinear interactions[1].

We focus on a specific class of such relationships, persistent infections, resulting from the pairing of a microbe and host that have survived the challenges of co-habitation. It is a phenotype defined by its success. Because these relationships are fundamentally different compared with either accidental or short-term co-evolved interactions, our goal is to clarify the key principles. The central concept we explore is that persistence represents the evolved selection for balancing host and microbial interests, resulting in an equilibrium that, by definition, is long-term but not necessarily forever stable. We hypothesize that maintenance of this equilibrium requires a series of evolved, nested equilibria to achieve the overall homeostasis.

The framework of such persistence is illustrated by examination of three bacterial species (*Helicobacter pylori*, *Salmonella typhi* and *Mycobacterium tuberculosis*) that are human-specific, despite causing well-recognized biological costs to their hosts[2–4]. These particular host–microbial interactions are representative of different classes of persistent infection (Fig. 1). We focus on bacterial infections of humans because of their importance and because of the knowledge already gained through their study[5]; however, the principles should be general to other microbes and other hosts.

Relationships between persistent microbes and their hosts span many spatial scales and timescales. At a microscopic timescale are the individual elements of both microbe and host reactive cell (immunocyte) populations, together with their intra-host evolution and interactions. The mesoscopic (physiological/ecological) scale involves population dynamics and interaction consequences for both host and transmission. At the macroscopic scale, host evolutionary changes occur[1].

We propose that microbial persistence represents a co-evolved series of nested equilibria, operating simultaneously on each of these multiple scales, to achieve an overall homeostasis. The composite equilibria of host and microbe may be considered as a 'holobiont'[6] (that is, organisms living together in symbiosis), regardless of whether there is mutualism[7]. Such relationships would resemble climax communities that have achieved stability under prevailing conditions. In the following sections, we consider elements critical

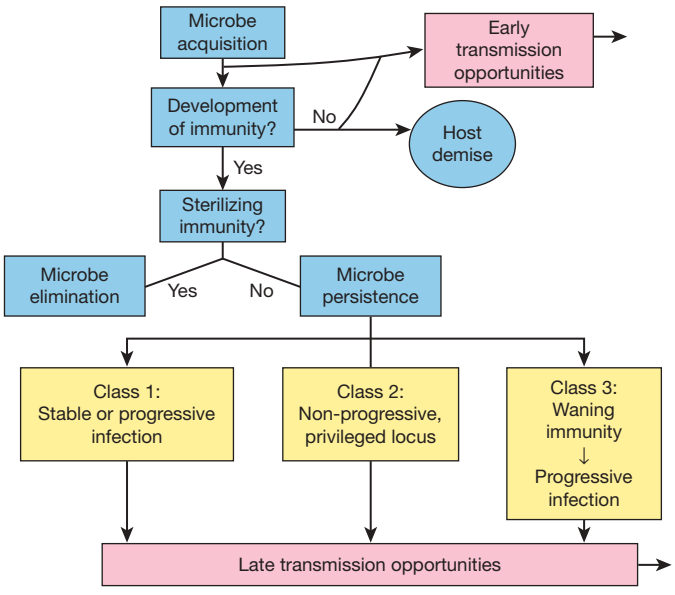


Figure 1 | Classes of microbial persistence. Because inter-host transmission is required for obligate host-associated microparasites, our model is organized according to transmission strategy. After microbial acquisition, there can be early transmission until effective immunity develops. For microbes able to resist immune elimination, late transmission may occur via progressive infection (class 1), non-progressive infection with carriage (class 2), or development of progressive infection in hosts with declining immunity (class 3). *H. pylori*, *S. typhi* and *M. tuberculosis* are representative human-associated microbes belonging to these three classes, respectively. Early and late transmission are biological trade-offs.

[1]Departments of Medicine and Microbiology, New York University School of Medicine, New York, New York 10016, USA. [2]Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.

Table 1 | Mechanisms used by persistent bacteria against host responses

Category	Principle	Example
Stealth	Intracellular location	*Chlamydia* species
	Sequestration (foreign body)	*Staphylococcus aureus*
	Molecular mimicry	*Escherichia coli* (K1 capsule)
	Low antigenicity (surface antigen)	*Treponema pallidum*
	Low expression of stimuli of innate responses	*Salmonella typhi* (Vi)
	Antigen masking	*Neisseria gonorrhoeae*
Variation	Surface-exposed antigens	*Bacteroides fragilis*
		Borrelia burgdorferi
Anti-defence	Antibody-absorbing	*Staphylococcus aureus* (protein A)
	IgA protease	*Haemophilus influenzae*
	Inhibition of phagocytosis	*Staphylococcus aureus*
	Resistance to phagocyte killing	*Salmonella*
	Disarming macrophages	*Yersinia* spp. (YOPs)
	Killing of macrophages	*Mycobacterium tuberculosis*
	Disarming T cells	*Helicobacter pylori* (VacA)

This Table is adapted from refs 52–54. YOPs, *Yersinia* outer proteins.

to the development of the equilibria, including generation of host immunity and its neutralization by persistent microbes (microscale); variation among populations of microbes and host cells (mesoscale); the parameters that affect inter-host microbial transmission (macroscale); and most critically, the types of rules governing the equilibria.

Immunity and microbial escape

Immunity, defined as the resistance of a host to the endogenous propagation of microbes, is mediated by innate or adaptive recognition[8]. Innate mechanisms are based on selection of hosts recognizing stereotypical structures, whereas adaptive immunity involves intra-host memory against encountered threats. Just as microbial populations evolved mechanisms to regulate group activities (for example, quorum sensing), processes evolved in hosts to regulate their immunocyte populations. In addition to upregulatory networks, regulatory T cells (T_R cells) have the ability to secrete chemical signals that limit T-helper 1 (T_H1) and T_H2 cellular responses[9–11]. Dedicated T_R cells[11,12] suppress auto-intolerance and limit the immunopathogenesis accompanying infections, probably selected by reducing tissue injury from infections[9]. The balance between T_R and T-effector cells affects infectious disease pathogenesis in individual hosts and at specific life-cycle stages[9].

By definition, persistent microbes have successful strategies to sufficiently thwart host responses to gain a niche. Many such microbial adaptations have been recognized, involving stealth, antigenic variation and anti-defence strategies (Table 1). Host responses may be narrow, with a single immune clone out-competing the others (immunodominance), or broad, in which multiple immune clones develop; efficient control of persistent infection correlates with narrow responses[13,14].

However, there is a balance between microbial immune evasion and maintaining growth fitness. The evolved microbial genome[15–17] reflects the tensions between these selective pressures[18,19]. For example, *H. pylori* both downregulates T-cell responses by secreting VacA[20], and upregulates mucosal signal transduction pathways by injecting into epithelial cells a protein (CagA) with tyrosine phosphorylation domains interacting with host cellular kinases and phosphatases[21–23]. Clonal variants within individual hosts differ in the number of phosphorylation domains, affecting interaction intensity[24]. The gene (*cagY*) encoding the injection system pilus protein possesses complex repetitive DNA regions that undergo intragenic recombination, creating antigenic variants[25]. Persistent *H. pylori* populations have been selected for their ability to manipulate T_R function[26].

Microbial transmission dynamics

For host-adapted microbes, transmission to new hosts is required. This concept is captured by the term R_0, which quantifies the transmission potential of a microparasite as the average number of secondary infections occurring when a single infectious host is introduced into a universally susceptible host population[27]. A simple way to define R_0 explicitly, on the basis of a standard model of epidemic transmission[28,29], is given by the equation:

$$R_0 = BN/(\alpha + b + v),$$

where BN is the transmission rate (a function of the population size, N), α is the rate of mortality owing to the microbe (a measure of virulence), b is the rate of mortality in the host population independent of the microbe (a measure of lifespan), and v is the rate at which hosts recover from the infection (a measure of immunity). In other formulations of R_0, although transmission rate is a function of B and N, the size of the population becomes more determinative[30]. When $R_0 > 1$, microbial transmission is sustained; when $R_0 < 1$, transmission goes to extinction.

The level of virulence is set by competition among microbes of the same species, because they always have the same host population number (N) at any given time. If the parameters of the R_0 equation are independent of one another, then the direction of evolution would be away from virulence towards commensalism, as selection would favour highly transmissible ($B \rightarrow$ large), persistent ($v \rightarrow 0$) commensals ($\alpha \rightarrow 0$) or symbionts ($\alpha \rightarrow -$large)[29]. However, if B and α are directly (and positively) related, then selection could favour some level of virulence ($\alpha > 0$) in the microbial population[27]. The effects of the introduction of myxoma virus to the rabbit population in Australia provides experimental support for this scenario[31].

There is further meaning to each of the terms of the R_0 equation. In much earlier times, when human populations were small[32], N was limiting, which selected against pathogens that had high mortality (α) (Table 2). With the rise of civilization[33,34], population growth, crowding and improved transportation, the number and proximity of susceptible hosts grew, which permitted more pathogenic organisms to flourish. Similarly, host variation (for example, immunodeficiencies increasing microbial number) affects B (the rate of transmission), thereby increasing R_0.

The basis for a host–microbe equilibrium model

Microbial success in a host requires the ability to grow and overcome the host's defences. The microbe must be able to access sufficient nutrients, overcome physical forces (such as the peristalsis of the gastrointestinal tract) and thwart innate or adaptive host defence molecules; these are host 'signals' to which the microbe must adapt. Conversely, microbial metabolites, toxins and anti-defence molecules, and physical adherence to host cells are microbial 'signals' to the host. The host-derived and microbial-derived signals may be either unlinked or linked. In the unlinked model, when the host wins, the microbe is eliminated, but if the microbe wins, the host dies. An

Table 2 | Ontogeny of microbe acquisition in human pre-history/history

Time-line of human history (yr BP)	Effective population size	Major source of microbial transmission	Nature of immunity	Example	Persistence
Most ancient (>50,000)	Isolated hunter-gatherers (<100)	Maternal/intrafamilial	Ineffective	*Bacteroides* species, *H. pylori*	Active
Intermediate (10,000–50,000)	Communicating hunter-gatherer groups (<100 to 10,000)	Long-term carriers	Containment but not elimination	*M. tuberculosis*, *S. typhi*, varicella-zoster virus	Latency
Recent (<10,000)	Large societies (>500,000)	Acutely ill persons	Life long	Measles	No
Very recent (<200)	>10 million	Acute infection	Serotype-specific	Pandemic influenza	No

alternative model, based on linked signals between microbe and host, implies selective pressure favouring co-evolved phenotypes[35,36], and is most applicable to persistent organisms (Fig. 2). In such a model, the host sequesters the bacterium into a discrete compartment (for example, the lumen of the gastrointestinal tract, the interior of a gallstone, the centre of a granuloma) that is surrounded by responding host cells that do not permit the microbe to extend into adjacent tissues[35–39]. A linkage between host and microbial signals and the achievement of persistence implies that equilibrium (homeostasis) has been reached.

The equilibrium model

To understand the principles permitting persistent equilibrium, we developed deterministic mathematical models[35,36]. Although we used *H. pylori* (Box 1) as the model organism, the underlying principles should be broadly generalizable. The essential feature of the model is that there must be both positive and negative feedback between the host and microbe; only with negative feedback can equilibrium (persistence) be achieved. The constructed model[36] encompasses five prototypic populations that are followed over time (Supplementary Information). There are two microbial subpopulations: bacteria that are free-living in the gastric mucus or are adherent to host cells. In a broader sense, these two populations also represent any two classes of bacterial cells that vary in the intensity of their host interactions. The model also defined a concentration of microbial effector molecules signalling the host, and a concentration of host-derived nutrients that benefit the microbes. Finally, the model included host immunity, governed by its response rate, ultimate capacity and the differential effect of the two microbial subpopulations with high or limited interaction. In this model, immunity limits microbial populations by restricting growth rates; immunity can be defined as lowering net microbial replication. By limiting replication, the autoregulatory network leads to either transient or persistent *H. pylori* colonization[36]. This model produced equilibrium solutions under a wide range of relevant biological variation. We propose that host status is also critical in determining the types of equilibrium reached with *S. typhi* (Box 2) and *M. tuberculosis* (Box 3).

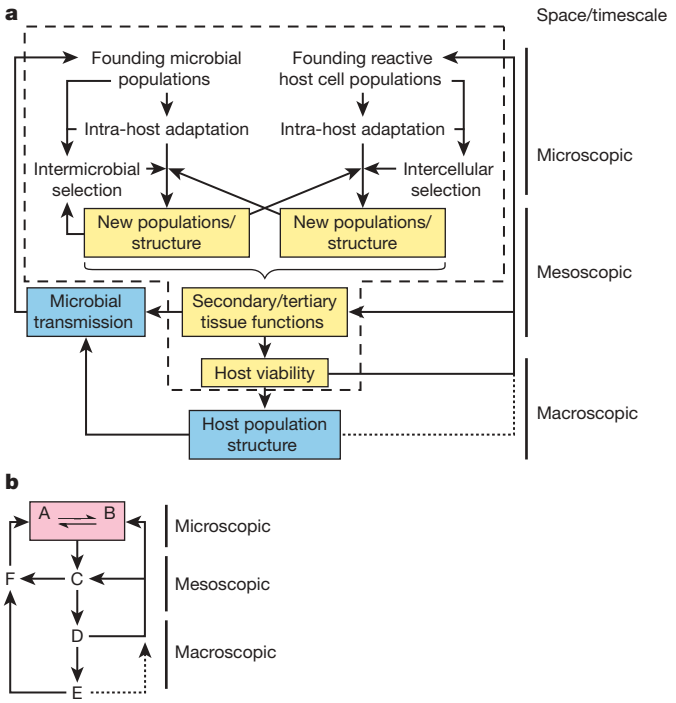


Figure 2 | A model for microbial persistence in metazoan hosts. **a**, Schematic with model elements. After founding microbes are acquired, new populations/population structure reflect intra-host adaptation, influenced by both intermicrobial selection (a product of microbial competition and cooperativity) and by the population of host reactive (immune) cells, which determine the resource space and structure. A parallel phenomenon describes the selection of host reactive cell populations. Events within the host are inside the dashed box. These microevolutionary events represent the first (microscopic) scale of the interaction (as adapted from ref. 1). For persisting organisms, these two interlocking phenomena have co-evolved and in their sum affect secondary tissue functions (for example, immune adjuvancy, hormone levels) that affect microbial transmission. These second scale (mesoscopic) interaction events influence host viability (for example, pathogens, through disease, or symbionts via resistance to pathogens or to famine). On the macroscopic (host evolutionary) interaction timescale, these events affect host population structure, which then governs microbial transmission and selection for host genotypes (shown by dotted lines). In this model, host population size and structure are important selectors for the types of microbes that can be successful. **b**, General schematic of the model. (See text for further discussion of elements A–F.)

Box 1 | Model 1: *Helicobacter pylori*

Helicobacter pylori, a Gram-negative bacterium that colonizes the human stomach as its sole environmental niche[23], has redundant (faecal–oral, oral–oral, vomitus–oral) transmission routes and notable biological success: (1) present in humans for >50,000 yr; (2) cosmopolitan (world-wide) in distribution; (3) nearly universal in all developing societies; (4) colonization essentially life-long; (5) dominant single species (>70% of clones) in the human stomach; and (6) multiple strains often colonizing the same host[55–60]. Once established in a host, *H. pylori* populations develop persistence, often life-long, with concomitant host responses. Although *H. pylori* enhances risk for lethal gastric cancer and peptic ulceration, these generally occur late in life[2]. Equilibrium involves stimulating host inflammation to provide nutrients[35–37]. Although *H. pylori* actively replicates in the human stomach for decades, persistence eventually may lead to progressive gastric atrophy[61], which reduces or eliminates its own colonization; thus, too much inflammation is maladaptive.

H. pylori has many characteristics favouring gastric colonization, outcompeting other microbes, and regulating and reducing specific immune responses[20,62–64]. Adaptive *H. pylori* genomic features include pathogenicity islands and numerous contingency genes[65], with variable expression based on binary switches[16]. On the basis of endogenous mutation and recombination[58], facilitated by natural competence for DNA uptake[66], *H. pylori* cells have high genomic plasticity, providing numerous phenotypic variants capable of colonizing diverse and changing niches[40]. Within-host dynamics between *H. pylori* competition[67] and cooperation[40] are an important tension (microscopic scale); success at this level provides one basis for the next stratum of equilibrium (Fig. 2), and the emergence of phenotypic variants[24,25] interacting with immunocytes[26] provides another basis.

At mesoscopic and macroscopic scales, the success of *H. pylori* in human populations reflects either low virulence or possible symbiosis early in life; essentially all of its negative consequences occur after reproductive age. Potential *H. pylori* benefits to young hosts include protection against diarrhoeal diseases and asthma, and metabolic regulation via gastric leptin and ghrelin[23,68–70].

Consistent with an ESS, host population structure influences *H. pylori* transmission. The family has been central[71], with early transmission opportunities from infected children[72], especially older siblings, later opportunities when girls become mothers (less so, but present from fathers), and later still, as an old age consequence of gastric atrophy and hypochlorhydria[73] (class 1 in Fig. 1). With low *H. pylori* virulence, there is little competition between early and late transmission (Fig. 1). Gastric *H. pylori* colonization is a prototypic ESS, as it effectively controls *H. pylori* cheaters, and essentially excludes all other bacteria from the stomach[56] for the bulk of an individual's life, well into the post-reproductive years, with probable early life benefits and little cost to hosts. However, the use of antibiotics in the 20th century may have eliminated an ESS that has existed since time immemorial.

Strain variation and the control of cheaters

The equilibrium model predicts that each microbial phenotypic variant develops different host interactions[35,36,40]. Bacterial variants often

Box 2 | Model 2: *Salmonella typhi*

Unlike all other known *Salmonella* species, *S. typhi* and the closely related *S. paratyphi A* are obligate pathogens of humans[74]. Because transmission is faecal–oral (often with food or water intermediates), the ability of *S. typhi* to enter the faecal stream by thwarting host phagocyte function[75,76] is critical. Under conditions of poor hygiene, *S. typhi* can infect large populations, but as those who survive natural infection (about 80%) develop permanent immunity, the pool of susceptible hosts is rapidly exhausted. However, some hosts become asymptomatic biliary carriers (for example, 'Typhoid Mary'), capable of life-long *S. typhi* transmission. Although the humoral and cellular immunity that develops[77,78] protects these hosts from disease, it is insufficient to sterilize the lumen of the gallbladder and biliary tract, especially when gallstones are present[79]; the stone becomes the segregated niche that enables life-long *S. typhi* carriage (class 2 in Fig. 1).

Genomic analysis of *S. typhi* has provided tools to understand its evolution[80]. The ancestral *S. typhi* haplotype arose after human migrations out of Africa (50,000 yr BP), but before the Neolithic period (10,000 yr BP)[74,81], when effective human population sizes were relatively small. Long persistence of individual haplotypes, neutral population structure and global transmission[74] are population correlates of a stable lifestyle with high biological success. *S. typhi* isolates show low genomic variation[74,80] consistent with stable immune interactions at the microscopic scale (Fig. 2). The effects of *S. typhi* on biliary tract function (mesoscopic scale) with stable carrier state development increase mean inter-host transmission time, allowing for spread within and between human groups when small populations were insufficient to sustain direct spread of acute infection. The long-term carrier keeps *S. typhi* extant in a population until new generations of susceptible hosts can be introduced to the organism (macroscopic level). As human populations grew along rivers[33,34], faecal contamination of portable water by a carrier could transmit *S. typhi* to distant downstream communities. Spread from carriers initiates epidemic cycles, seeding populations for generations to come.

Box 3 | Model 3: *Mycobacterium tuberculosis*

Mycobacterium tuberculosis, the cause of tuberculosis, is usually transmitted from diseased hosts via cough-borne aerosols[82]. Once acquired by the respiratory route, *M. tuberculosis* establishes a pulmonary parenchymal focus, but in most hosts, life-long latency develops, with low likelihood for disease. This raises the question of how latency is established.

The essential site for *M. tuberculosis* persistence is within the granuloma, a complex structure of bacteria and host multinucleated cells and infected macrophages, encircled by both activated and non-activated macrophages and T cells. The granuloma has a central caseous necrotic core that harbours mycobacteria and dead host cells[83,84]. In this environment, host cells and mycobacteria can interact for the host's lifetime[85], with bacterial replication[86,87] but controlled growth. Mathematical models[88–93] of the conditions favouring latency have defined two bacterial subpopulations (intracellular and extracellular) with distinctive growth rates and signals to host cells, and different macrophage-response states and T-cell and cytokine contributions[92]. An equilibrium maintains low bacterial levels and controls tissue damage, based on macrophage activation and control[83,84,93]. This is the locus for the microscopic scale of the general model (Fig. 2).

The models predict that disease reactivation occurs when the granulomas no longer effectively control extracellular bacterial growth; when intracellular management predominates, local tissue damage and bacterial dissemination are reduced. The models also predict that the signalling that occurs between host cells and their intracellular bacteria facilitate granuloma maintenance, and that the slow mycobacterial growth rate favours latency[85,88–91,93]. Thus, *M. tuberculosis* has evolved slow growth rates and the ability to survive inside macrophages, while hosts who minimize tissue damage from potentially over-zealous immune responses have been selected[75,83,85].

However, with ageing or other immunodeficiencies, infection within the granuloma is no longer suppressed, net bacterial growth accelerates, and disease occurs (mesoscopic scale)[82,94]. This 'reactivation' form of tuberculosis is most common, creating new opportunities for transmission via coughing, often decades after the organism was acquired (class 3 in Fig. 1). With both early and late transmission possibilities, *M. tuberculosis* can skip generations of human hosts, an effective strategy for host populations of small size (macroscopic scale). The global population structure of *M. tuberculosis*, defined by phylogeographic lineages associated with sympatric human populations[95], provides evidence for its co-evolution with humans, as does the disproportion of allopatric tuberculosis cases involving compromised hosts[96]. That host population characteristics determine the extent of clinical tuberculosis in a community[97] is consistent with the remarkable genomic conservation of *M. tuberculosis*[17,98]. As predicted, certain 'cheater' events, such as utilization of forbidden sites (for example, development of tuberculous meningitis), cause host demise without transmission; strains exhibiting such phenotypes would be selected against.

arise through mutation, intragenomic recombination, or horizontal gene transfer[40,41]. When hosts harbour more than one strain simultaneously, these compete, but often also cooperate (through genetic exchange and specialized function)[42,43]. The model indicates that for competitors to persist, each must occupy an exclusive niche, or face eventual elimination[35,36]. An implicit limitation of an equilibrium model is the emergence of individuals ('cheaters') that break the rules to their own advantage[44].

Game theory provides solutions for how nature can resolve this dilemma. A cheater may be defined as a player that changes strategy unilaterally. A Nash equilibrium is a strategy profile in a game with $\geq$2 players in which none can gain by changing strategy unilaterally[45]. A subset of the Nash equilibrium is the evolutionarily stable strategy (ESS)[46,47], which when present in a population resists invasion by a competing alternative strategy. We propose that co-evolved persistent microbe–host systems have developed ESSs, which preclude cheater success.

What boundaries would ensure ESS maintenance? Because the persistence model is based on linked regulation of host and microbial signals, a cheater is a variant signalling for resources but not halting its growth when the resources are provided, as the equilibrium requires. One solution to this problem is that penalties for transgression have evolved in the ESS that ultimately lower cheater fitness. Penalties can involve crossing thresholds to induce new host responses. A host response whereby bacterial growth triggers new innate or adaptive responses with subsequent amplification would be effective, as any growth advantage for the cheater would be temporary and local. Because a novel mutant can escape the specific immunity directed towards a predominant strain, ecosystem stability might favour microbes with low mutation rates[48]. However, the penalty mechanism, affecting all strains of the microbe including cheaters, does not permit mutational escape. Regulatory T cells are a class of immunocyte that could closely modulate host responses to microbial perturbations[9–11], but multiple mechanisms exist (Supplementary Information).

A general model of microbial persistence in hosts

Despite the enormous microbial variation that exists, our prior mathematical modelling and examination of three cases of microbial persistence (Boxes 1–3) indicate that a general hypothesis for persistence in metazoan hosts can be developed. In complex ecosystems, such as within humans, the model depends on a series of evolved equilibrium relationships, nested in one another and interconnected, and operating simultaneously over three different biological scales. The model proposed represents an ESS, and has six major components (identified as A–F; Fig. 2).

Element A represents the microbial populations persisting in a particular tissue or host compartment (Fig. 2a). The composition and structure of the population is based on the founding populations, the intra-host generation of variation, the selection imposed by the competing (and cooperating) microbes, and the selection

imposed by the host. The composition and population structure of the reactive host cells involved in innate and adaptive immunity (element B) is based on principles parallel to those governing the microbial cells (founders, variants generated, selective pressure from competing/cooperating cells) and the selection exerted by the persisting microbes. Thus, the two populations (A and B) are interdependent, and exist in a linked dynamic equilibrium.

The nature of this primary (microscale) host–microbial equilibrium shapes tissue function (element C, mesoscale), which ultimately affects both host viability (element D, macroscale) and microbial transmission (element F, mesoscale). Pathogenic microbes damage tissue, leading to coughing, vomiting or diarrhoea, favouring their own transmission. Conversely, the tissue effects of symbionts are protective (for example, metabolic or immune), selecting for the hosts that carry them.

The (negative or positive) effects on host viability select for host genes in elements B and C, influencing population structure (element E), which through extinction vortices also affects the host gene pool (elements B and C). The host population structure affects microbial transmission (element F), influencing the founding microbial populations in new hosts; small host population size selects against virulence, and short lifespans select against late-transmitting microbes.

This is a dynamic model of co-evolved hosts and microbes (Fig. 2b), requiring multiple scales, flexible across a range of conditions, and useful for understanding both symbionts and pathogens. In reality, there is no fixed distinction between the two; their biological behaviour is defined by their ecological context.

Discussion

Microbial transmission—central to the maintenance of persistent host-adapted infections—is considered as being vertical across generations or horizontal across populations. Typically, indigenous (commensal) organisms are transmitted vertically from mother to child, whereas pathogens are transmitted horizontally. However, there are intermediate cases[49], because an individual is more likely to cough on family members than on strangers, and the microbes transmitted from mother to offspring may be affected by her environmental exposures. R_0 dynamics can be affected by mixed vertical and horizontal transmission, as well as by demographic changes, such as number of births per woman.

Transfer of a microbe to a host genetically related to the previous host occurs with vertical, but not necessarily horizontal, transmission; as pandemic infections become more frequent in the modern world, horizontal transmission has an enhanced role. Microbial genomes are plastic, with extensive intra-host variation[37,38]; strains partly adapted to a new host owing to passage through a genetically related previous host may yield different outcomes than strains from unrelated persons[50].

As predicted by the R_0 equation, with small effective population sizes the long hunter-gatherer stage of human evolution was a bottleneck for highly virulent human pathogens. Small population sizes selected for symbionts or for pathogens that could be transmitted decades after infecting a host, after new susceptible individuals had been introduced into the population via births (Table 2). In contrast, high-virulence pathogens would have been driven to extinction by the demise of their isolated host populations. However, with the larger effective population sizes that have developed since the rise of agriculture[33,34], more virulent pathogens have been appearing. Our rapidly changing human context, including widespread immunodeficiencies and jet travel, is continuing to alter the selection for human-adapted microbes.

For example, the proportion of hosts newly infected with *M. tuberculosis* who develop progressive tuberculosis and become immediately infectious, who reactivate the infection late, or who never reactivate, is dependent on the immunocompetence of the host population. Host characteristics unevenly distributed across the population, including malnutrition and HIV infection, affect the proportions of individuals in each compartment and thus, the transmission profiles. Similarly, because tuberculosis reactivation rates are age-dependent, general improvements in health that lead to increased proportions of elderly persons in the population affect outcomes. Conversely, reactivation of lethal infections tends to keep overall host lifespan under close regulation. Nevertheless, for microbes like *M. tuberculosis*, there is also a cost to latency, because competing mortality limits transmission. As HIV has become more common, there has been selection towards progressive primary tuberculosis.

As illustrated by *M. tuberculosis*, the evolution of a persistent parasite that uses latency as part of its transmission strategy integrates the transmission rates for all stages in the host life cycle, keeping net $R_0 > 1$. The balance between early and late opportunities for transmission is context specific, dependent on host variables including effective population size, age structure, distribution of immunocompetence and previous selection for resistance. Similarly for symbionts, context matters. A microbe that induces iron deficiency may be symbiotic in regions where malaria is holoendemic[51], but without malaria may decrease host fitness. Because context is all-important in evolution, the multiple scales on which persistent parasitic and symbiotic infections operate provide substrate for the dynamic solutions that unfold.

We propose a new model based on ESSs, a subset of Nash equilibria, to explain the common features of microbial persistence in their human hosts. That the model was consistent with the observed biology of three bacteria (*H. pylori*, *S. typhi* and *M. tuberculosis*) with highly dissimilar genomic and lifestyle features supports its generalizability. Importantly, the model applies to both pathogens and commensals, and can be used to understand the direction of virulence as the context of human ecology changes.

1. Law, R. & Dieckmann, U. Symbiosis through exploitation and the merger of lineages in evolution. *Proc R. Soc. Lond. B* **265,** 1245–1253 (1998).
2. Peek, R. M. & Blaser, M. J. *Helicobacter pylori* and gastrointestinal tract adenocarcinomas. *Nature Rev. Cancer* **2,** 28–37 (2002).
3. Hornick, R. B. *et al.* Typhoid fever: pathogenesis and immunologic control. *N. Engl. J. Med.* **283,** 686–691 (1970).
4. Glickman, M. & Jacobs, W. Microbial pathogenesis of *Mycobacterium tuberculosis*: dawn of a discipline. *Cell* **104,** 477–485 (2003).
5. Rosebury, T. *Microorganisms Indigenous to Man* 1–8 (McGraw Hill, New York, 1962).
6. Margulis, L. *Symbiosis in Cell Evolution* 2nd edn 163 (W.H Freeman, New York, 1993).
7. Lewis, D. H. Symbiosis and mutualism: crisp concepts and soggy semantics. In *The Biology of Mutualism: Ecology and Evolution* (ed. Boucher, D. H.) 29–39 (Croom Helm, London, 1985).
8. Medzhitov, R. Recognition of microorganisms and activation of the immune response. *Nature* doi:10.1038/nature06246 (this issue).
9. Belkaid, Y. & Rouse, B. T. Natural regulatory T cells in infectious disease. *Nature Immunol.* **6,** 353–360 (2005).
10. Fontenot, J. D. & Rudensky, A. A well adapted regulatory contrivance: regulatory T cell development and the Forkhead family transcription factor Foxp3. *Nature Immunol.* **6,** 331–337 (2005).
11. Sakaguchi, S. *et al.* $Foxp3^+CD25^+CD4^+$ natural regulatory T cells in dominant self-tolerance and autoimmune disease. *Immunol. Rev.* **212,** 8–27 (2006).
12. Zheng, Y. & Rudensky, A. Y. Foxp3 in control of the regulatory T cell lineage. *Nature Immunol.* **8,** 457–462 (2007).
13. Wodarz, D. & Nowak, M. A. CD8 memory immunodominance and antigenic escape. *Eur. J. Immunol.* **30,** 2704–2712 (2000).
14. Wodarz, D. & Nowak, M. A. Correlates of CTL-mediated virus control; implications for immunosuppresive infections and their treatment. *Phil. Trans. R. Soc. Lond. B* **355,** 1059–1070 (2000).
15. Aras, R. A., Kang, J., Tschumi, A., Harasaki, Y. & Blaser, M. J. Extensive repetitive DNA facilitates prokaryotic genome plasticity. *Proc. Natl Acad. Sci. USA* **100,** 13579–13584 (2003).
16. Saunders, N. J., Peden, J. F., Hood, D. W. & Moxon, E. R. Simple sequence repeats in the *Helicobacter pylori* genome. *Mol. Microbiol.* **27,** 1091–1098 (1998).
17. Fleischmann, R. D. *et al.* Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184,** 5479–5490 (2002).
18. Bonhoeffer, S. & Nowak, M. Intra-host versus inter-host selection: viral strategies of immune function impairment. *Proc. Natl Acad. Sci. USA* **91,** 8062–8066 (1994).
19. Nowak, M. & May, R. Superinfection and the evolution of parasite virulence. *Proc. Biol. Sci.* **225,** 81–89 (1994).

20. Gebert, B., Fischer, W., Weiss, E., Hoffmann, R. & Haas, R. *Helicobacter pylori* vacuolating cytotoxin inhibits T lymphocyte activation. *Science* **301,** 1099–1102 (2003).
21. Odenbreit, S. *et al.* Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. *Science* **287,** 1497–1500 (2000).
22. Yokoyama, K. *et al.* Functional antagonism between *Helicobacter pylori* CagA and vacuolating toxin VacA in control of the NFAT signaling pathway in gastric epithelial cells. *Proc. Natl Acad. Sci. USA* **102,** 9661–9666 (2005).
23. Blaser, M. J. & Atherton, J. *Helicobacter pylori* persistence: biology and disease. *J. Clin. Invest.* **113,** 321–333 (2004).
24. Aras, R. A. *et al.* Natural variation in populations of persistently colonizing bacteria affect human host cell phenotype. *J. Infect. Dis.* **188,** 486–496 (2003).
25. Aras, R. A. *et al.* Plasticity of repetitive DNA sequences within a bacterial (type IV) secretion system component. *J. Exp. Med.* **198,** 1349–1360 (2003).
26. Lundgren, A., Suri-Payer, E., Enarsson, K., Svennerholm, A. M. & Lundin, B. S. *Helicobacter pylori*-specific $CD4^+CD25^{high}$ regulatory T cells suppress memory T-cell responses to *H. pylori* in infected individuals. *Infect. Immun.* **71,** 1755–1762 (2003).
27. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* 17–19 (Oxford Univ. Press, Oxford, 1991).
28. Anderson, R. M. & May, R. M. Co-evolution of hosts and parasites. *Parasitology* **85,** 411–426 (1982).
29. Levin, B. R. The evolution and maintenance of virulence in microparasites. *Emerg. Infect. Dis.* **2,** 93–102 (1996).
30. Dietz, K. Overall population patterns in the transmission cycle of infectious agents. In *Population Biology of Infectious Diseases* (eds Anderson, R. & May, R.) 87–102 (Springer, Berlin, 1982).
31. Fenner, F. & Ratcliffe, F. N. *Myxomatosis* (Cambridge Univ. Press, Cambridge, 1965).
32. Barnard, A. J. (ed.) *Hunter-gatherers in History, Archeology and Anthropology* 278 (Berg, Oxford, 2004).
33. Bellwood, P. *First Farmers: the Origins of Agricultural Societies* 360 (Blackwell, Oxford, 2004).
34. Smith, B. D. *The Emergence of Agriculture* 231 (Scientific American Library, New York, 1995).
35. Kirschner, D. E. & Blaser, M. J. The dynamics of *Helicobacter pylori* infection of the human stomach. *J. Theor. Biol.* **176,** 281–290 (1995).
36. Blaser, M. J. & Kirschner, D. Dynamics of *Helicobacter pylori* colonization in relation to the host response. *Proc. Natl Acad. Sci. USA* **96,** 8359–8364 (1999).
37. Falk, P. G. *et al.* Theoretical and experimental approaches for studying factors that define the relationship between *Helicobacter pylori* and its host. *Trends Microbiol.* **8,** 321–329 (2000).
38. Kirschner, D. & Marino, S. *Mycobacterium tuberculosis* as viewed through a computer. *Trends Microbiol.* **13,** 206–211 (2005).
39. Bledzka-Sarek, M. & El Skurnik, M. How to outwit the enemy: dendritic cells face *Salmonella*. *APMIS* **144,** 589–600 (2006).
40. Levine, S. M. *et al.* Plastic cells and populations: DNA substrate characteristics in *Helicobacter pylori* transformation define a flexible but conservative system for genomic variation. *FASEB J.* (in the press).
41. Krinos, C. M. *et al.* Extensive surface diversity of a commensal microorganism by multiple DNA inversions. *Nature* **414,** 555–558 (2001).
42. Smith, J. The social evolution of bacterial pathogenesis. *Proc. R. Soc. Lond. B* **261,** 61–69 (2001).
43. Fiegna, F., Yu, Y.-T. N., Kadam, S. V. & Velicer, G. J. Evolution of an obligate social cheater to a superior cooperator. *Nature* **441,** 310–314 (2006).
44. Neumann, J. V. & Morgenstern, O. *Theory of Games and Economic Behavior* (Princeton Univ. Press, Princeton, New Jersey, 1944).
45. Nash, J. Non-cooperative games. *Ann. Math.* **54,** 286–295 (1951).
46. Smith, J. M. & Price, G. R. The logic of animal conflict. *Nature* **246,** 15–18 (1973).
47. Smith, J. M. Evolution and the theory of games. *Am. Sci.* **64,** 41–45 (1976).
48. Nowak, M. & May, R. *Mathematical Principles of Immunology and Virology* (Oxford Univ. Press, New York, 2000).
49. Hope-Simpson, R. E. Infectiousness of communicable diseases in the household. *Lancet* **2,** 549–554 (1952).
50. Blaser, M. J., Nomura, A., Lee, J., Stemmerman, G. N. & Perez-Perez, G. I. Early life family structure and microbially-induced cancer risk. *PLoS Med.* **4,** e7 (2007).
51. Dominguez-Bello, M. G. & Blaser, M. J. Are iron-scavenging parasites protective against malaria? *J. Infect. Dis.* **191,** 646 (2005).
52. Mims, C. A., Dimmock, N. J., Nash, A. & Stephen, J. Microbial strategies in relation to the immune response. In *Mim's Pathogenesis of Infectious Diseases* 4th edn 168–196 (Academic, San Diego, 1995).
53. Monack, D. M., Mueller, A. & Falkow, S. Persistent bacterial infections: the interface of the pathogen and the host immune system. *Nature Rev. Microbiol.* **2,** 747–765 (2004).
54. Young, D., Hussell, T. & Dougan, G. Chronic bacterial infections: living with unwanted guests. *Nature Immunol.* **3,** 1026–1032 (2002).
55. Linz, B. *et al.* An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445,** 915–918 (2007).
56. Bik, E. M. *et al.* Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl Acad. Sci. USA* **103,** 732–737 (2006).
57. Blaser, M. J. Who are we? Indigenous microbes and the ecology of human diseases. *EMBO Rep.* **7,** 956–960 (2006).
58. Falush, D. *et al.* Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: Estimates of clock rates, recombination size, and minimal age. *Proc. Natl Acad. Sci. USA* **98,** 15056–15061 (2001).
59. Wirth, H. P. *et al.* Host Lewis phenotype-dependent *Helicobacter pylori* Lewis antigen expression in rhesus monkeys. *FASEB J.* **20,** 1534–1536 (2006).
60. Ghose, C., Perez-Perez, G. I., van Doorn, L. J., Dominguez-Bello, M. G. & Blaser, M. J. High frequency of gastric colonization with multiple *Helicobacter pylori* strains in Venezuelan subjects. *J. Clin. Microbiol.* **43,** 2635–2641 (2005).
61. Kuipers, E. J. *et al.* Long-term sequelae of *Helicobacter pylori* gastritis. *Lancet* **345,** 1525–1528 (1995).
62. McGowan, C. C., Necheva, A. S., Forsyth, M. H., Cover, T. L. & Blaser, M. J. Promoter analysis of *Helicobacter pylori* genes whose expression is enhanced at low pH. *Mol. Microbiol.* **48,** 1225–1239 (2003).
63. Kim, S.-Y., Lee, Y.-C., Kim, H. K. & Blaser, M. J. *Helicobacter pylori* CagA transfection of gastric epithelial cells induces interleukin-8. *Cell. Microbiol.* **8,** 97–106 (2006).
64. O'Brien, D. P. *et al.* The role of decay-accelerating factor as a receptor for *Helicobacter pylori* and a mediator of gastric inflammation. *J. Biol. Chem.* **281,** 13317–13323 (2006).
65. Tomb, J. F. *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388,** 539–547 (1997).
66. Kang, J. M. & Blaser, M. J. Bacterial populations as perfect gases: genomic diversity and diversification tensions in *Helicobacter pylori*. *Nature Rev. Microbiol.* **4,** 826–836 (2006).
67. Webb, G. F. & Blaser, M. J. Dynamics of bacterial phenotype selection in a colonized host. *Proc. Natl Acad. Sci. USA* **99,** 3135–3140 (2002).
68. Putsep, K., Branden, C. I., Boman, H. G. & Nomark, S. Antibacterial peptide from *H. pylori*. *Nature* **398,** 671–672 (1999).
69. Chen, Y. & Blaser, M. J. Inverse associations of *Helicobacter pylori* with asthma and allergies. *Arch. Intern. Med.* **167,** 821–827 (2007).
70. Nwokolo, C. U., Freshwater, D. A., O'Hare, P. & Randeva, H. S. Plasma ghrelin following cure of *Helicobacter pylori*. *Gut* **52,** 637–640 (2003).
71. Raymond, J. *et al.* Genetic and transmission analysis of *Helicobacter pylori* strains within a family. *Emerg. Infect. Dis.* **10,** 1816–1821 (2004).
72. Parsonnet, J., Shmuely, H. & Haggerty, T. D. Fecal and oral shedding of *Helicobacter pylori* from healthy, infected adults. *J. Am. Med. Assoc.* **282,** 2240–2245 (1999).
73. Fox, J. G. *et al.* Role of gastric pH in isolation of *Helicobacter mustelae* from the feces of ferrets. *Gastroenterology* **104,** 86–92 (1993).
74. Roumagnac, P. *et al.* Evolutionary history of *Salmonella typhi*. *Science* **314,** 1301–1304 (2006).
75. Schwan, W. R. *et al.* Differential bacterial survival, replication, and apoptosis-inducing ability of *Salmonella* serovars within human and murine macrophages. *Infect. Immun.* **68,** 1005–1013 (2000).
76. Robbins, J. D. & Robbins, J. B. Reexamination of the protective role of the capsular polysaccharide (Vi antigen) of *Salmonella typhi*. *J. Infect. Dis.* **150,** 436–449 (1984).
77. Espersen, F. *et al.* Humoral and cellular immunity in typhoid and paratyphoid carrier state, investigated by means of quantitative immunoelectrophoresis and *in vitro* stimulation of blood lymphocytes. *Acta Pathol. Microbiol. Immunol. Scand.* **90,** 293–299 (1982).
78. Faucher, S. P. *et al.* Transcriptome of *Salmonella enterica* serovar Typhi within macrophages revealed through the selective capture of transcribed sequences. *Proc. Natl Acad. Sci. USA* **103,** 1906–1911 (2006).
79. Sinnott, C. R. & Teall, A. J. Persistent gallbladder carriage of *Salmonella typhi*. *Lancet* **1,** 976 (1987).
80. Parkhill, J. *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413,** 848–852 (2001).
81. Kidgell, C. *et al. Salmonella typhi*, the causative agent of typhoid fever, is approximately 50,000 years old. *Infect. Genet. Evol.* **2,** 39–45 (2002).
82. Clark-Curtiss, J. E. & Haydel, S. E. Molecular genetics of *Mycobacterium tuberculosis* pathogenesis. *Annu. Rev. Microbiol.* **57,** 517–549 (2003).
83. Lazarevic, V., Nolt, D. & Flynn, J. L. Long-term control of *Mycobacterium tuberculosis* infection is mediated by dynamic immune responses. *J. Immunol.* **175,** 1107–1117 (2005).
84. Lin, P. L. *et al.* Early events in *Mycobacterium tuberculosis* infection in cynomolgus macaques. *Infect. Immun.* **74,** 3790–3803 (2006).
85. Marino, S., Pawr, S., Reinhart, T. A., Flynn, J. L. & Kirschner, D. E. Dendritic cell trafficking and antigen presentation in the human immune response to *Mycobacterium tuberculosis*. *J. Immunol.* **173,** 494–506 (2004).
86. Capuano, S. V. *et al.* Experimental *Mycobacterium tuberculosis* infection of cynomolgus macaques closely resembles the various manifestations of human *M. tuberculosis* infection. *Infect. Immun.* **71,** 5831–5844 (2003).
87. Munoz-Elias, E. J. *et al.* Replication dynamics of *Mycobacterium tuberculosis* in chronically infected mice. *Infect. Immun.* **73,** 546–551 (2005).
88. Wigginton, J. E. & Kirschner, D. A model to predict cell-mediated immune regulatory mechanisms during human infection with *Mycobacterium tuberculosis*. *J. Immunol.* **166,** 1951–1976 (2001).
89. Marino, S. & Kirschner, D. The human immune response to *Mycobacterium tuberculosis* in the lung and lymph node. *J. Theor. Biol.* **227,** 463–486 (2004).
90. Gammack, D., Doering, C. & Kirschner, D. Macrophage response to *Mycobacterium tuberculosis* infection. *J. Math. Biol.* **48,** 218–242 (2003).

91. Ganguli, S., Gammack, D. & Kirschner, D. A metapopulation model of granuloma formation in the lung during infection with *M. tuberculosis*. *Math. Biosci. Engin.* **22,** 535–560 (2005).
92. Sud, D., Bigbee, C., Flynn, J. L. & Kirschner, D. E. Contribution of $CD8^+$ T cells to control of *Mycobacterium tuberculosis* infection. *J. Immunol.* **176,** 4296–4314 (2006).
93. Segovia-Juarez, J., Ganguli, S. & Kirschner, D. Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent based model. *J. Theor. Biol.* **231,** 357–376 (2004).
94. Blower, S. M. *et al.* The intrinsic transmission dynamics of tuberculosis epidemics. *Nature Med.* **1,** 815–821 (1995).
95. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. & Small, P. M. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc. Natl Acad. Sci. USA* **1001,** 4871–4876 (2004).
96. Gagneux, S. *et al.* Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **103,** 2869–2873 (2006).
97. Murphy, B. M., Singer, B. H., Anderson, S. & Kirschner, D. Comparing epidemic tuberculosis in demographically distinct heterogeneous populations. *Math. Biosci.* **180,** 161–185 (2005).
98. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393,** 537–544 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the NIH, the Ellison Medical Foundation and by the Diane Belfer Program for Human Microbial Ecology in Health and Disease. We thank D. Krakauer and Y. Iwasa for discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence should be addressed to M.B. (martin.blaser@med.nyu.edu).

ARTICLES

A second generation human haplotype map of over 3.1 million SNPs

The International HapMap Consortium*

We describe the Phase II HapMap, which characterizes over 3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals from four geographically diverse populations and includes 25–35% of common SNP variation in the populations surveyed. The map is estimated to capture untyped common variation with an average maximum r^2 of between 0.9 and 0.96 depending on population. We demonstrate that the current generation of commercial genome-wide genotyping products captures common Phase II SNPs with an average maximum r^2 of up to 0.8 in African and up to 0.95 in non-African populations, and that potential gains in power in association studies can be obtained through imputation. These data also reveal novel aspects of the structure of linkage disequilibrium. We show that 10–30% of pairs of individuals within a population share at least one region of extended genetic identity arising from recent ancestry and that up to 1% of all common variants are untaggable, primarily because they lie within recombination hotspots. We show that recombination rates vary systematically around genes and between genes of different function. Finally, we demonstrate increased differentiation at non-synonymous, compared to synonymous, SNPs, resulting from systematic differences in the strength or efficacy of natural selection between populations.

Advances made possible by the Phase I haplotype map

The International HapMap Project was launched in 2002 with the aim of providing a public resource to accelerate medical genetic research. The objective was to genotype at least one common SNP every 5 kilobases (kb) across the euchromatic portion of the genome in 270 individuals from four geographically diverse populations[1,2]: 30 mother–father–adult child trios from the Yoruba in Ibadan, Nigeria (abbreviated YRI); 30 trios of northern and western European ancestry living in Utah from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (CEU); 45 unrelated Han Chinese individuals in Beijing, China (CHB); and 45 unrelated Japanese individuals in Tokyo, Japan (JPT). The YRI samples and the CEU samples each form an analysis panel; the CHB and JPT samples together form an analysis panel. Approximately 1.3 million SNPs were genotyped in Phase I of the project, and a description of this resource was published in 2005 (ref. 3).

The initial HapMap Project data had a central role in the development of methods for the design and analysis of genome-wide association studies. These advances, alongside the release of commercial platforms for performing economically viable genome-wide genotyping, have led to a new phase in human medical genetics. Already, large-scale studies have identified novel loci involved in multiple complex diseases[4,5]. In addition, the HapMap data have led to novel insights into the distribution and causes of recombination hotspots[3,6], the prevalence of structural variation[7,8] and the identity of genes that have experienced recent adaptive evolution[3,9]. Because the HapMap cell lines are publicly available, many groups have been able to integrate their own experimental data with the genome-wide SNP data to gain new insight into copy-number variation[10], the relationship between classical human leukocyte antigen (HLA) types and SNP variation[11], and heritable influences on gene expression[12–14]. The ability to combine genome-wide data on such diverse aspects of genetic variation with molecular phenotypes collected in the same samples provides a powerful framework to study the connection of DNA sequence to function.

*Lists of participants and affiliations appear at the end of the paper.

In Phase II of the HapMap Project, a further 2.1 million SNPs were successfully genotyped on the same individuals. The resulting HapMap has an SNP density of approximately one per kilobase and is estimated to contain approximately 25–35% of all the 9–10 million common SNPs (minor allele frequency (MAF) ≥ 0.05) in the assembled human genome (that is, excluding gaps in the reference sequence alignment; see Supplementary Text 1), although this number shows extensive local variation. This paper describes the Phase II resource, its implications for genome-wide association studies and additional insights into the fine-scale structure of linkage disequilibrium, recombination and natural selection.

Construction of the Phase II HapMap

Most of the additional genotype data for the Phase II HapMap were obtained using the Perlegen amplicon-based platform[15]. Briefly, this platform uses custom oligonucleotide arrays to type SNPs in DNA segmentally amplified via long-range polymerase chain reaction (PCR). Genotyping was attempted at 4,373,926 distinct SNPs, which corresponds, with exceptions (see Methods), to nearly all SNPs in dbSNP release 122 for which an assay could be designed. Additional submissions were included from the Affymetrix GeneChip Mapping Array 500K set, the Illumina HumanHap100 and HumanHap300 SNP assays, a set of ~11,000 non-synonymous SNPs genotyped by Affymetrix (ParAllele) and a set of ~4,500 SNPs within the extended major histocompatibility complex (MHC)[11]. Genotype submissions were subjected to the same quality control (QC) filters as described previously (see Methods) and mapped to NCBI build 35 (University of California at Santa Cruz (UCSC) hg17) of the human genome. The re-mapping of SNPs from Phase I of the project identified 21,177 SNPs that had an ambiguous position or some other feature indicative of low reliability; these are not included in the filtered Phase II data release. All genotype data are available from the HapMap Data Coordination Center (http://www.hapmap.org) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP); analyses described in this paper refer to release 21a. Three data sets are available: 'redundant unfiltered'

contains all genotype submissions, 'redundant filtered' contains all submissions that pass QC, and 'non-redundant filtered' contains a single QC+ submission for each SNP in each analysis panel.

The QC filters remove SNPs showing gross errors. However, it is also important to understand the magnitude and structure of more subtle genotyping errors among SNPs that pass QC. We therefore carried out a series of analyses to assess the influence of the long-range PCR amplicon structure on genotyping error, the concordance rates between genotype calls from different genotyping platforms and between those platforms and re-sequencing assays, as well as the rates of false monomorphism and mis-mapping of SNPs (see Supplementary Text 2, Supplementary Figs 1–3 and Supplementary Tables 1–4). We estimate that the average per genotype accuracy is at least 99.5%. However, there are higher rates of missing data and genotype discrepancies at non-reference alleles, with some clustering of errors resulting from the amplicon design and a few incorrectly mapped SNPs.

Table 1 shows the numbers of SNPs attempted and converted to QC+ SNPs in each analysis panel (Supplementary Table 5 shows a breakdown by each major submission). Haplotypes and missing data were estimated for each analysis panel separately using both trio information and statistical methods based on the coalescent model (see Methods). To enable cross-population comparisons, a consensus data set was created consisting of 3,107,620 SNPs that were QC+ in all analysis panels and polymorphic in at least one analysis panel. The equivalent figure from Phase I was 931,340 SNPs. Unless stated otherwise, all analyses have been carried out on the consensus data set. An additional set of haplotypes was created for those SNPs in the consensus where a putative ancestral state could be assigned by comparison of the human alleles to the orthologous position in the chimpanzee and rhesus macaque genomes.

The variation in SNP density within the Phase II HapMap is shown in Fig. 1. On average there are 1.14 genotyped polymorphic SNPs per kilobase (average spacing is 875 base pairs (bp)) and 98.6% of the assembled genome is within 5 kb of the nearest polymorphic SNP. Still, there is heterogeneity in genotyped SNP density at both broad (Fig. 1a) and fine (Fig. 1b) scales. Furthermore, there are systematic changes in genotyped SNP density around genomic features including genes (Fig. 1c).

The Phase II HapMap differs from the Phase I HapMap not only in SNP spacing, but also in minor allele frequency distribution and patterns of linkage disequilibrium (Supplementary Fig. 4). Because the criteria for choosing additional SNPs did not include consideration of SNP spacing or preferential selection for high MAF, the SNPs added in Phase II are, on average, more clustered and have lower MAF than the Phase I SNPs. Because MAF predictably influences the distribution of linkage disequilibrium statistics, the average r^2 at a given physical distance is typically lower in Phase II than in Phase I; conversely, the $|D'|$ statistic is typically higher (data not shown). One notable consequence is that the Phase II HapMap includes a better representation of rare variation than the Phase I HapMap.

The increased resolution provided by Phase II of the project is illustrated in Fig. 2. Broadly, an additional SNP added to a region shows one of three patterns. First, it may be very similar in distribution to SNPs present in Phase I. Second, it may provide detailed resolution of haplotype structure (for example, a group of chromosomes with identical local haplotypes in Phase I can be shown in Phase II to carry

Table 1 | Summary of Phase II HapMap data (release 21)

Phase	SNP categories	Analysis panel		
		YRI	CEU	CHB+JPT
I	Assays submitted	1,304,199	1,344,616	1,306,125
	Passed QC	1,177,312 (90%)	1,217,902 (91%)	1,187,800 (91%)
	Did not pass QC	126,887 (10%)	126,714 (9%)	118,325 (9%)
	>20% missing	82,463 (65%)	95,684 (76%)	78,323 (66%)
	>1 duplicate inconsistent	6,049 (5%)	5,126 (4%)	9,242 (8%)
	>1 mendelian error	18,916 (15%)	11,310 (9%)	N/A
	<0.001 Hardy–Weinberg *P* -value	10,265 (8%)	8,922 (7%)	13,722 (12%)
	Other failures	19,345 (15%)	13,858 (11%)	20,674 (17%)
II	Assays submitted	5,044,989	5,044,996	5,043,775
	Passed QC	3,150,433 (62%)	3,204,709 (64%)	3,244,897 (64%)
	Did not pass QC	1,894,556 (38%)	1,840,287 (36%)	1,798,878 (36%)
	>20% missing	1,419,000 (75%)	1,398,166 (76%)	1,403,543 (78%)
	>1 duplicate inconsistent	0 (0%)	0 (0%)	6,617 (0%)
	>1 mendelian error	172,339 (9%)	127,923 (7%)	N/A
	<0.001 Hardy–Weinberg *P* -value	96,231 (5%)	82,268 (4%)	108,880 (6%)
	Other failures	334,511 (18%)	337,906 (18%)	340,370 (19%)
Overall	Assays submitted	6,349,188	6,389,612	6,349,900
	Passed QC	4,327,745 (68%)	4,422,611 (69%)	4,432,697 (70%)
	Did not pass QC	2,021,443 (32%)	1,967,001 (31%)	1,917,203 (30%)
	>20% missing	1,501,463 (74%)	1,493,850 (76%)	1,481,866 (77%)
	>1 duplicate inconsistent	6,049 (0%)	5,126 (0%)	15,859 (1%)
	>1 mendelian error	191,255 (9%)	139,233 (7%)	N/A
	<0.001 Hardy–Weinberg *P* -value	106,496 (5%)	91,190 (5%)	122,602 (6%)
	Other failures	353,856 (18%)	351,764 (18%)	361,044 (19%)
	Non-redundant (unique) SNPs	3,796,934	3,868,157	3,890,416
	Monomorphic	861,299 (23%)	1,246,183 (32%)	1,410,152 (36%)
	Polymorphic	2,935,635 (77%)	2,621,974 (68%)	2,480,264 (64%)

	SNP categories	All analysis panels
	Unique QC-passed SNPs	4,000,107
	Passed in one analysis panel	88,140 (2%)
	Passed in two analysis panels	268,534 (7%)
	Passed in three analysis panels (QC+3)	3,643,433 (91%)
	QC+3 and monomorphic across three analysis panels	535,813
	QC+3 and polymorphic in at least one analysis panel	3,107,620
	QC+3 and polymorphic in all three analysis panels	2,006,352
	QC+3 and MAF ≥ 0.05 in at least one of three analysis panels	2,819,322

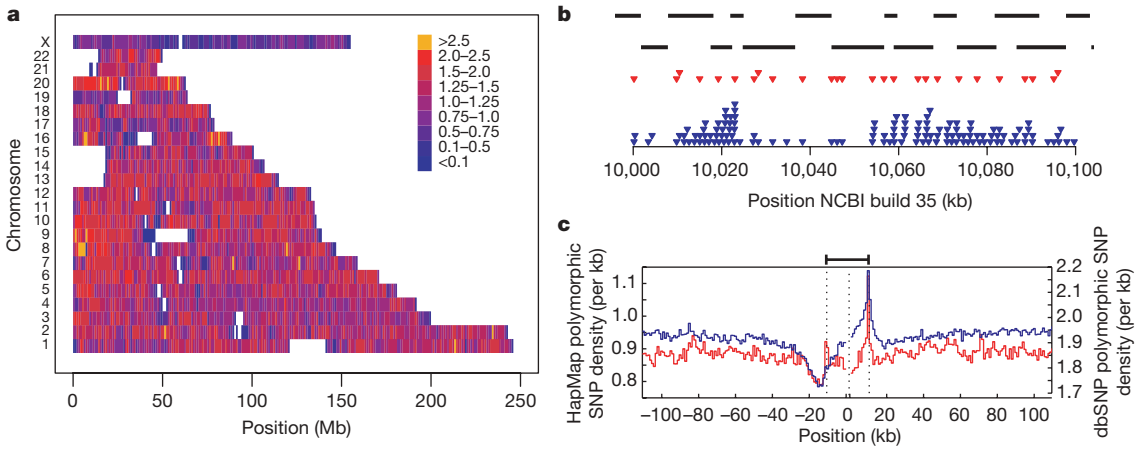


Figure 1 | SNP density in the Phase II HapMap. a, SNP density across the genome. Colours indicate the number of polymorphic SNPs per kb in the consensus data set. Gaps in the assembly are shown as white. **b**, Example of the fine-scale structure of SNP density for a 100-kb region on chromosome 17 showing Perlegen amplicons (black bars), polymorphic Phase I SNPs in the consensus data set (red triangles) and polymorphic Phase II SNPs in the consensus data set (blue triangles). Note the relatively even spacing of Phase I SNPs. **c**, The distribution of polymorphic SNPs in the consensus Phase II HapMap data (blue line and left-hand axis) around coding regions. Also shown is the density of SNPs in dbSNP release 125 around genes (red line and right-hand axis). Values were calculated separately 5′ from the coding start site (the left dotted line) and 3′ from the coding end site (right dotted line) and were joined at the median midpoint position of the coding unit (central dotted line).

multiple related haplotypes). Third, the novel SNP (or group of added SNPs) may reveal previously missed recombinant haplotypes. The extent to which each type of event occurs varies among populations and chromosomal regions. The greatest gains in resolution, in terms of identifying new recombinant haplotypes and haplotype groupings, occur in YRI. Consequently, the Phase II HapMap provides increased resolution in the estimated fine-scale genetic map and improved power to detect and localize recombination hotspots (Fig. 2b).

The use of the Phase II HapMap in association studies

The increased SNP density of the Phase II HapMap has already been extensively exploited in genome-wide studies of disease association. In this section, we quantify the gain in resolution and outline how the HapMap data can be used to improve the power of association studies.

Improved coverage of common variation. We previously predicted that the vast majority of common SNPs would be correlated to Phase II HapMap SNPs by extrapolation from the ten HapMap ENCODE regions[3]. Using the actual Phase II marker spacing and frequency distributions (Table 2), we repeated the simulations and estimate that Phase II HapMap marker sets capture the overwhelming majority of all common variants at high r^2. For common variants (MAF $\geq$ 0.05) the mean maximum r^2 of any SNP to a typed one is 0.90 in YRI, 0.96 in CEU and 0.95 in CHB+JPT. The impact of the

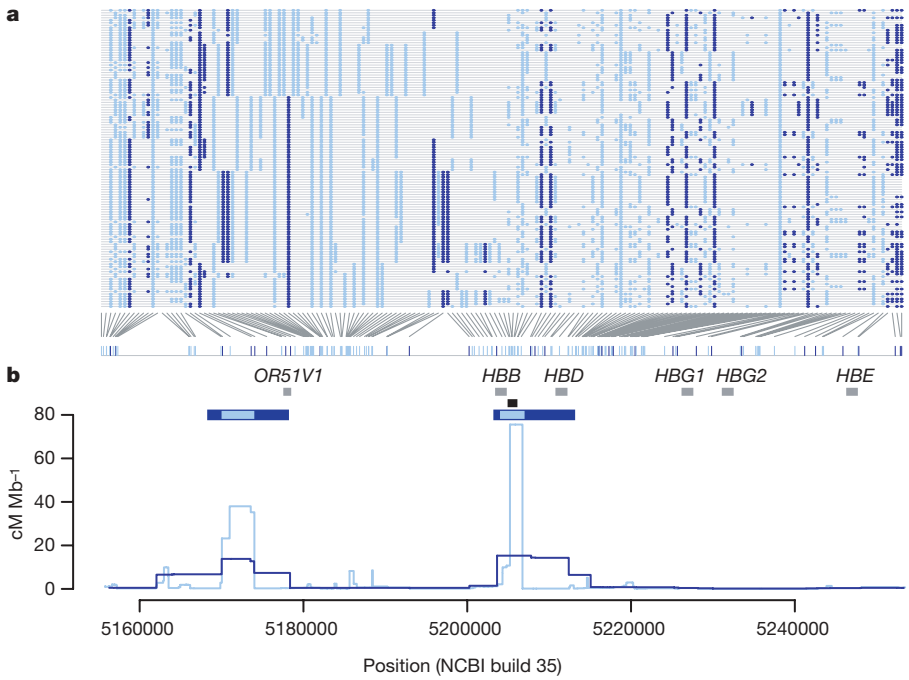


Figure 2 | Haplotype structure and recombination rate estimates from the Phase II HapMap. a, Haplotypes from YRI in a 100 kb region around the β-globin (*HBB*) gene. SNPs typed in Phase I are shown in dark blue. Additional SNPs in the Phase II HapMap are shown in light blue. Only SNPs for which the derived allele can be unambiguously identified by parsimony (by comparison with an outgroup sequence) are shown (89% of SNPs in the region); the derived allele is shown in colour. **b**, Recombination rates (lines) and the location of hotspots (horizontal blue bars) estimated for the same region from the Phase I (dark blue) and Phase II HapMap (light blue) data. Also shown are the location of genes within the region (grey bars) and the location of the experimentally verified recombination hotspot[57,58] at the 5′ end of the *HBB* gene (black bar).

Table 2 | Estimated coverage of the Phase II HapMap in the ten HapMap ENCODE regions

Panel	MAF bin	Phase I HapMap[3]		Phase II HapMap			
				Pairwise linkage disequilibrium		Additional 2-SNP tests	
		$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2
YRI	≥0.05	45	0.67	82	0.90	87	0.93
	<0.05			61	0.76	62	0.78
	0.05–0.10			81	0.89	81	0.89
	0.10–0.25			90	0.94	90	0.95
	0.25–0.50			87	0.93	92	0.96
CEU	≥0.05	74	0.85	93	0.96	95	0.97
	<0.05			70	0.79	72	0.81
	0.05–0.10			87	0.92	88	0.93
	0.10–0.25			94	0.96	95	0.97
	0.25–0.50			95	0.97	97	0.98
CHB+JPT	≥0.05	72	0.83	92	0.95	95	0.97
	<0.05			65	0.74	65	0.74
	0.05–0.10			81	0.89	82	0.89
	0.10–0.25			90	0.94	90	0.95
	0.25–0.50			94	0.96	97	0.98

2-SNP tests, linkage disequilibrium to haplotypes formed from two nearby SNPs.

Table 3 | Number of tag SNPs required to capture common (MAF ≥ 0.05) Phase II SNPs

Threshold	YRI	CEU	CHB+JPT
$r^2 \geq 0.5$	627,458	290,969	277,831
$r^2 \geq 0.8$	1,093,422	552,853	520,111
$r^2 = 1.0$	1,616,739	1,024,665	1,078,959

increased density of the Phase II HapMap is most notable in YRI (in the Phase I HapMap the mean maximum r^2 was 0.67). Similar results are found if a threshold of $r^2 \geq 0.8$ is used to determine whether an SNP is captured (Table 2). As expected, very common SNPs with MAF > 0.25 are captured extremely well (mean maximum r^2 of 0.93 in YRI to 0.97 in CEU), whereas rarer SNPs with MAF < 0.05 are less well covered (mean maximum r^2 of 0.74 in CHB+JPT to 0.76 in YRI). The latter figure is probably an overestimate because it is based on lower frequency SNPs discovered via re-sequencing 48 HapMap individuals, and does not include a much larger number of very rare SNPs. We also assessed the increase in coverage provided by using two-SNP haplotypes as proxies for SNPs that are poorly captured by single SNPs[16] (Table 2). These two-SNP haplotypes lead to a modest increase in mean maximum r^2 of 0.01 to 0.03 across all allele frequencies. However, in some regions, particularly where marker density is low, gains from multi-marker and imputation approaches in practical situations can be substantial (see below).

Currently, the Phase II HapMap provides the most complete available resource for selecting tag SNPs genome-wide. Using a simple pairwise tagging approach, we find that 1.09 million SNPs are required to capture all common Phase II SNPs with $r^2 \geq 0.8$ in YRI, with slightly more than 500,000 required in CEU and CHB+JPT (Table 3). These numbers are approximately twice those required to capture SNPs in the Phase I HapMap (which has one-third as many SNPs). The number of SNPs required to achieve perfect tagging ($r^2 = 1.0$) in each analysis panel is almost double that required to achieve the $r^2 \geq 0.8$ threshold. It becomes increasingly expensive to improve the coverage afforded by tags from the Phase I and, now, the Phase II HapMap, because additional tag SNPs are unlikely to capture large groups of additional SNPs.

Phase II HapMap and genome-wide association studies. Although the efficient choice of tag SNPs is one use of the Phase II HapMap, for most disease studies the tag SNPs genotyped will be primarily determined by the choice of a commercial platform for the experiment[17,18]. Using Phase II data, we estimated the coverage of several available products on which genome-wide association studies are already underway (Table 4). Similar to earlier estimates[17,18], these products typically perform well in CEU and CHB+JPT, and some also perform well in YRI. For example, arrays of approximately 500,000 SNPs capture 68–88% (depending on selection method) of all HapMap Phase II variation with $r^2 \geq 0.8$ in CEU. SNPs that are not included in the Phase II HapMap will be covered more poorly because most genotyping products were designed using HapMap data.

HapMap data have several additional roles in the analysis of disease-association studies using fixed marker sets. For example, the high-quality haplotype information within the Phase II HapMap can be used to aid the phasing of genotype data from new samples because additional haplotypes are likely to be locally very similar to at least one haplotype in the Phase II data. By a similar argument, missing genotypes can potentially be inferred through comparison to the Phase II haplotypes. Genotypes may be missing either because of genotyping failure or because the SNP was not assayed within the experiment. Therefore, the HapMap haplotypes provide a way of *in silico* genotyping Phase II SNPs that were not included in the experiment.

Although there is no clear consensus yet about the role of SNP imputation in the analysis of genome-wide association studies, high imputation accuracy can be achieved using model-based methods[19–23] and can lead to an increase in power[23,24]. To illustrate the possibilities, in the 500-kb HapMap ENCODE region on 8q24.11 (Supplementary Fig. 5) we evaluated imputation of Phase II SNPs from the Affymetrix GeneChip 500K array. To do this, we used a

Table 4 | Estimated coverage of commercially available fixed marker arrays

Platform*	YRI		CEU		CHB+JPT	
	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2	$r^2 \geq 0.8$ (%)	Mean maximum r^2
Affymetrix GeneChip 500K	46	0.66	68	0.81	67	0.80
Affymetrix SNP Array 6.0	66	0.80	82	0.90	81	0.89
Illumina HumanHap300	33	0.56	77	0.86	63	0.78
Illumina HumanHap550	55	0.73	88	0.92	83	0.89
Illumina HumanHap650Y	66	0.80	89	0.93	84	0.90
Perlegen 600K	47	0.68	92	0.94	84	0.90

* Assuming all SNPs on the product are informative and pass QC; in practice these numbers are overestimates.

leave-one-out procedure to assess the accuracy of genotype prediction in the YRI. For SNPs with MAF ≥ 0.2, the average maximum r^2 to a typed SNP in the region is 0.59 compared to an average genotype prediction r^2 of 0.86. Furthermore, whereas 44% of such SNPs in the region have no single-marker proxy with $r^2 \geq 0.5$, fewer than 6% of the SNPs have a genotype imputation accuracy of $r^2 < 0.5$, establishing that accurate imputation can be achieved even in the population where linkage disequilibrium is the weakest.

New insights into linkage disequilibrium structure

The paradigm underlying association studies is that linkage disequilibrium can be used to capture associations between markers and nearby untyped SNPs. However, the Phase II HapMap has revealed several properties of linkage disequilibrium that illustrate the full complexity of empirical patterns of genetic variation. Two striking features are the long-range similarity among haplotypes, and SNPs that show almost no linkage disequilibrium with any other SNP.

The extent of recent common ancestry and segmental sharing. A simplified view of linkage disequilibrium is that genetic variation is organized in relatively short stretches of strong linkage disequilibrium (haplotype blocks), each containing only a few common haplotypes and separated by recombination hotspots across which little association remains[25]. Although this view has heuristic value, if chromosomes share a recent common ancestor then similarity between chromosomes can extend over considerable genetic distance and span multiple recombination hotspots[26]. The extent of such recent ancestry in the four populations surveyed here has not been characterized previously. Therefore we identified stretches of identity between pairs of chromosomes, both within and across individuals, reflecting autozygosity and identity-by-descent (IBD) (Fig. 3a). After first checking for stratification within each analysis panel (see Supplementary Text 3; none was found for YRI, CEU and JPT, and only small stratification was found for CHB), we calculated genome-wide probabilities of sharing 0, 1 or 2 chromosomes identical by descent for each pair of individuals (see Supplementary Text 4). In addition to identifying a few close relationships (as reported in HapMap Phase I[3]), we estimate that, on average, any two individuals from the same population share approximately 0.5% of their genome through recent IBD (Table 5). Using a hidden Markov model approach[27] (see Supplementary Text 5), we searched for such shared segments over 1-megabase (Mb) long and containing at least 50 SNPs, after first pruning the list of SNPs to remove local linkage disequilibrium. We find that 10–30% of pairs in each analysis panel share regions of extended identity resulting from sharing a common ancestor within 10–100 generations. These regions typically span hundreds of SNPs and can extend over tens of megabases (Table 5).

Similarly, extended stretches of homozygosity are indicative of recent inbreeding within populations[28,29]. Although short runs of homozygosity are commonplace, covering up to one-third of the genome and showing population differences reflective of ancient linkage disequilibrium patterns (Table 5 and Fig. 3b), very long homozygous runs exist that are clearly distinct from this process. Including two JPT individuals who have unusually high levels of homozygosity (NA18987 and NA18992) and one CEU individual (NA12874), we identified 79 homozygous regions over 3 Mb in 51 individuals, with many segments extending over 10 Mb (Supplementary Tables 7 and 8). Segments intersecting with suspected deletions were first removed from the analysis (Supplementary Text 6).

In studies of rare mendelian diseases, the extended haplotype sharing surrounding recent mutations, usually with a frequency of much less than 1%, has been exploited to great advantage through homozygosity mapping[30,31] and haplotype sharing[32] methods. In studies of common disease, extended haplotype sharing among patients potentially offers a route for identifying rare variants (MAF in the range of 1–5%) of high penetrance[33,34], which tend to be poorly captured through single-marker association with genome-wide arrays. To illustrate the idea, we identified SNPs where only two copies of the minor allele are present (referred to as '2-SNPs'), which have minor allele frequencies of 1–2%. We find that these are enriched approximately sevenfold (Table 5) among regions of IBD identified by the hidden Markov model approach. Notably, identification of IBD regions can be performed with the same genome-wide SNP data being

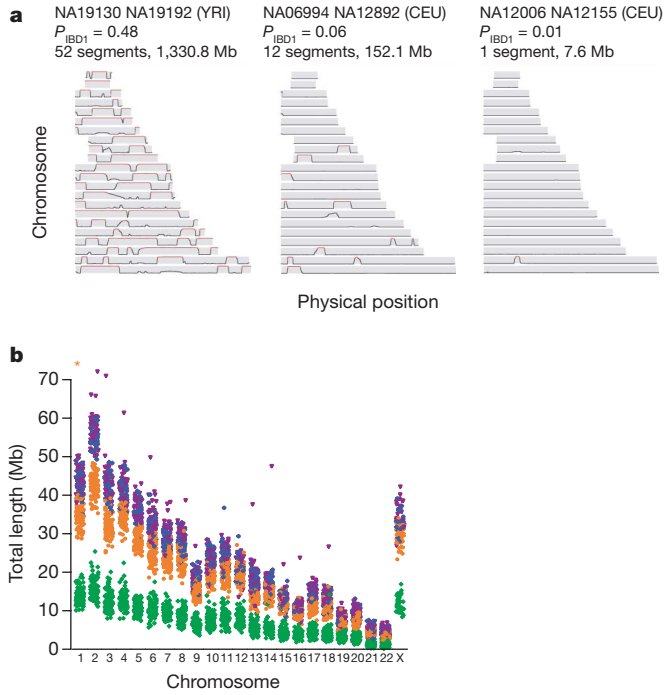


Figure 3 | The extent of recent co-ancestry among HapMap individuals. **a**, Three pairs of individuals with varying levels of identity-by-descent (IBD) sharing illustrate the continuum between very close and very distant relatedness and its relation to segmental sharing. The three pairs are: high sharing (NA19130 and NA19192 from YRI; previously identified as second-degree relatives[3]), moderate sharing (NA06994 and NA12892 from CEU) and low sharing (NA12006 and NA12155 from CEU). Along each chromosome, the probability of sharing at least one chromosome IBD is plotted, based on the HMM method described in Supplementary Text 5. Red sections indicate regions called as segments: in general, the proportion of the genome in segments is similar to each pair's estimated global relatedness. **b**, The extent of homozygosity on each chromosome for each individual in each analysis panel. Excludes segments <106 kb and chromosome X in males. Asterisk, NA12874, length = 107 Mb. YRI, green; CEU, orange; CHB, blue; JPT, magenta.

Table 5 | Relatedness, extended segmental sharing and homozygosity

Property	YRI	CEU	CHB	JPT
Number of pairs included	1,767	1,708	990	861
Mean identity by state (IBS) (%)	81.9	83.7	85.0	85.1
Mean identity by descent (IBD) (%)	0.04	0.34	0.36	0.42
Number of pairs with >1% IBD (%)	8.8	20.4	21.1	29.7
Number of pairs with one or more segment (%)	195 (11.0)	350 (20.5)	135 (13.6)	216 (25.1)
Total number of segments	250	427	146	273
Total distance spanned (Mb)	1,416	2,336	704	1,301
Mean segment length (Mb)	5.7	5.5	4.8	4.8
Maximum segment length (Mb)	51.7	56.2	15.0	25.3
Maximum segment length (Mb) (including close relatives)	141.4	128.5	N/A	N/A
Total number of 2-SNPs	6,219	9,220	8,174	8,750
Number of 2-SNPs in segments	109	162	116	132
2-SNP fold increase	6.7	7.3	7.6	7.0
Number of homozygous segments ($\times 10^3$)*	0.9	2.2	2.6	2.6
SNPs in homozygous segments ($\times 10^5$)	1.6	4.2	5.3	5.4
Total length of homozygous segments (Mb)	160	410	510	520

2-SNP, SNPs where only two copies of the minor allele are present.
* Homozygous segments >106 kb.

collected in large-scale association studies, making haplotype-sharing approaches an attractive and complementary analysis to standard SNP association tests, with the potential to identify rare variants associated with complex disease.

The distribution and causes of untaggable SNPs. Despite the SNP density of the Phase II HapMap, there are high-frequency SNPs for which no tag can be identified. Among high-frequency SNPs (MAF $\geq$ 0.2), we marked as untaggable SNPs to which no other SNP within 100 kb has an r^2 value of at least 0.2. In Phase II, approximately 0.5–1.0% of all high-frequency SNPs are untaggable and the proportion in YRI is approximately twice as high as in the other panels. Similar proportions are observed across the ten HapMap ENCODE regions.

To identify factors influencing the location of untaggable SNPs we considered their distribution relative to segmental duplications, repeat sequence, CpG dinucleotide density, regions of low SNP density, unusual allele frequency distribution, linkage disequilibrium patterns and recombination hotspots. We find no evidence for an enrichment of untaggable SNPs in segmental duplications or repeat sequence, as would be expected from mis-mapping of SNPs (2% and 35% of common SNPs lie in segmental duplications and repeat sequence, respectively, compared to 1.8% and 29%, respectively, of untaggable SNPs). Untaggable SNPs are slightly enriched in CpG islands (0.37% of common SNPs are in CpG islands compared to 1.4% of untaggable SNPs) and have slightly reduced MAF (Fig. 4). Most notably, untaggable SNPs are strongly enriched in regions of low linkage disequilibrium, particularly in recombination hotspots. To test whether these untaggable SNPs are themselves responsible for the identification of recombination hotspots, we eliminated them from 100 randomly chosen recombination hotspots and reassessed the evidence for a local peak in recombination. In all cases we still find evidence for a considerable increase in local recombination rate.

Over 50% of all untaggable SNPs lie within 1 kb of the centre of a detected recombination hotspot and over 90% are within 5 kb. Because only 3–4% of all SNPs lie within 1 kb from the centre of a detected recombination hotspot (16% are within 5 kb), this constitutes a marked enrichment and implies that at least 10% of all SNPs within 1 kb of hotspots are untaggable. The implication for association mapping is that when a region of interest contains a known hotspot it may be prudent to perform additional sequencing within the hotspot. Many of the variants identified in this manner will be untaggable SNPs that should be genotyped directly in association studies. From a biological perspective, the proximity of untaggable SNPs to the centre of hotspots suggests that they may lie within gene conversion tracts associated with the repair of double-strand breaks. Double-strand breaks are thought to resolve as crossover events only 5–25% of the time[35]. Consequently, SNPs lying near the centre of a hotspot are liable to be included within gene conversion tracts and will experience much higher effective recombination rates than predicted from crossover rates alone.

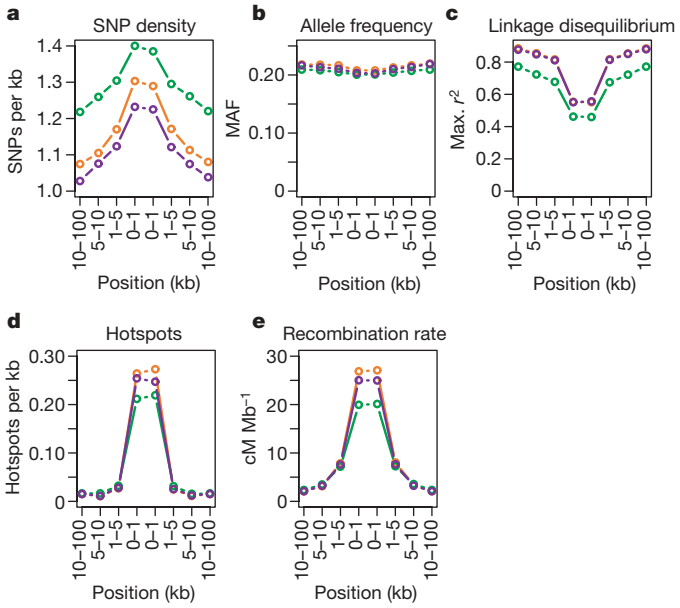


Figure 4 | Properties of untaggable SNPs. **a**–**e**, Properties of the genomic regions surrounding untaggable SNPs in terms of: **a**, the density of polymorphic SNPs within the consensus data set; **b**, mean minor allele frequency of polymorphic SNPs; **c**, maximum r^2 of SNPs to any others in the Phase II data; **d**, the density of estimated recombination hotspots (defined from hotspot centres); and **e**, the estimated mean recombination rate. YRI, green; CEU, orange; CHB+JPT, purple.

The distribution of recombination

In the Phase II HapMap we identified 32,996 recombination hotspots[3,6,36] (an increase of over 50% from Phase I) of which 68% localized to a region of $\leq$5 kb. The median map distance induced by a hotspot is 0.043 cM (or one crossover per 2,300 meioses) and the hottest identified, on chromosome 20, is 1.2 cM (one crossover per 80 meioses). Hotspots account for approximately 60% of recombination in the human genome and about 6% of sequence (Supplementary Fig. 6). We do not find marked differences among chromosomes in the concentration of recombination in hotspots, which implies that obligate differences in recombination among chromosomes of different size result from differences in hotspot density and intensity[6].

The increased number of well-defined hotspots allows us to understand better the influence of genomic features on the distribution of recombination. Previous work identified specific DNA motifs that influence hotspot location[6,37] as well as additional influences of local sequence context including the location of genes[6] and base composition[38]. The Phase II HapMap provides the resolution to separate these influences. Figure 5a shows the distribution of recombination, hotspot motifs and base composition around genes. Within the transcribed region of genes there is a marked decrease in the estimated recombination rate. However, 5′ of the transcription start site is a peak in recombination rate with a corresponding local increase in the density of hotspot motifs. This region also shows a marked increase in G+C content, reflecting the presence of CpG islands in promoter regions. There is also an asymmetry in recombination rate across genes, with recombination rates 3′ of transcribed regions being elevated (as are motif density and G+C content) compared to regions 5′ of genes. Studies in yeast have previously suggested an association between promoter regions and recombination hotspots[39]. Our results suggest a significant, although weak, relationship between promoters and recombination in humans. Nevertheless, the vast majority of hotspots in the human genome are not in gene promoters. The association may reflect a general association between regions of accessible chromatin and crossover activity.

Systematic differences in recombination rate by gene class. Previous work has demonstrated differences in the magnitude of linkage disequilibrium, as measured at a megabase scale, among genes associated with different functions[3,40]. Using the fine-scale genetic map estimated from the Phase II HapMap data we can quantify local increases in recombination rate associated with genes of different function using the Panther gene ontology annotation[41]. Average recombination rates vary more than sixfold among such gene classes (Fig. 5b), with defence and immunity genes showing the highest rates (1.9 cM Mb^{-1}) and chaperones showing the lowest rates (0.3 cM Mb^{-1}). Gene functions associated with cell surfaces and external functions tend to show higher recombination rates (immunity, cell adhesion, extracellular matrix, ion channels, signalling) whereas those with lower recombination rates are typically internal to cells (chaperones, ligase, isomerase, synthase). Controlling for systematic differences between gene classes in base composition and gene clustering, the differences between groups remain significant.

We also find that the density of hotspot-associated DNA motifs varies systematically among gene classes and that variation in motif density explains over 50% of the variance in recombination rate among gene functions (Supplementary Fig. 7).

These results pose interesting evolutionary questions. Because recombination involves DNA damage through double-strand breaks, hotspots may be selected against in some highly conserved parts of the genome. In regions exposed to recurrent selection (for example, from changes in environment or pathogen pressure) it is plausible that recombination may be selected for. However, because the fine-scale structure of recombination seems to evolve rapidly[42,43] it will be important to learn whether patterns of recombination rate heterogeneity among molecular functions are conserved between species.

Natural selection

The Phase I HapMap data have been used to identify genomic regions that show evidence for the influence of adaptive evolution[3,9], primarily through extended haplotype structure indicative of recent positive selection. Using two established approaches[9,44], we identified approximately 200 regions with evidence of recent positive selection from the Phase II HapMap (Supplementary Table 9). These regions include many established cases of selection, such as the genes *HBB* and *LCT*, the HLA region, and an inversion on chromosome 17. Many other regions have been previously identified in HapMap Phase I including *LARGE*, *SYT1* and *SULT1C2* (previously called *SULT1C1*). A detailed description of the findings from the Phase II HapMap is published elsewhere[45].

The Phase II HapMap also provides new insights into the forces acting on SNPs in coding regions. Effort was made to genotype as many known or putative non-synonymous SNPs as possible. Of the 56,789 non-synonymous SNPs identified in dbSNP release 125, attempts were made to genotype 36,777, which resulted in 17,427 that are QC+ in all three analysis panels and polymorphic. We selected only those SNPs for which ancestral allele information was available (approximately 90%). For comparison, we used patterns of variation at synonymous SNPs. As previously reported[46,47], non-synonymous SNPs show an increase in frequency of rare variants and a slight decrease of common variants compared to synonymous SNPs, compatible with widespread purifying selection against non-synonymous mutations (Fig. 6a). In contrast, we find no excess of high-frequency derived non-synonymous mutations, as might be expected if positive selection were widespread.

Natural selection also influences the extent to which allele frequencies differ between populations, not only through local selective pressures that drive alleles to different frequencies[48,49], but also through local variation in the strength of purifying selection. We compared the distribution of population differentiation (as measured by F_{ST}, the proportion of total variation in allele frequency that is due to differences between populations) at non-synonymous SNPs and synonymous SNPs matched for allele frequency (Fig. 6b). We find a systematic bias for non-synonymous SNPs to show stronger differentiation than synonymous SNPs. Among SNPs showing high levels of differentiation there is a strong tendency for the derived allele to be at higher frequency in non-YRI populations. Among SNPs with $F_{ST} > 0.5$ between CEU and YRI, in 79% and 75% of non-synonymous and synonymous variants, respectively, the derived allele is more common in CEU. Although this difference between non-synonymous and synonymous SNPs is not significant, among the eight exonic SNPs with $F_{ST} > 0.95$, all are non-synonymous. We see no such bias towards increased MAF in CEU at high-differentiation SNPs, indicating that SNP ascertainment is unlikely to explain the difference. Rather, this effect can largely be explained by more genetic drift in the non-African populations, as confirmed by simulations (data not shown). In addition, reduced selection against deleterious mutations and local adaptation within non-African populations will both act to increase the frequency of derived variants in non-African populations.

To assess the evidence for widespread local adaptation influencing non-synonymous mutations we considered the distribution of integrated extended haplotype homozygosity (iEHH) statistics[9,44] (Fig. 6c). We find no evidence for systematic differences between non-synonymous and synonymous SNPs, suggesting that local adaptation does not explain their higher differentiation. Although hitch-hiking effects will tend to obscure differences between selected

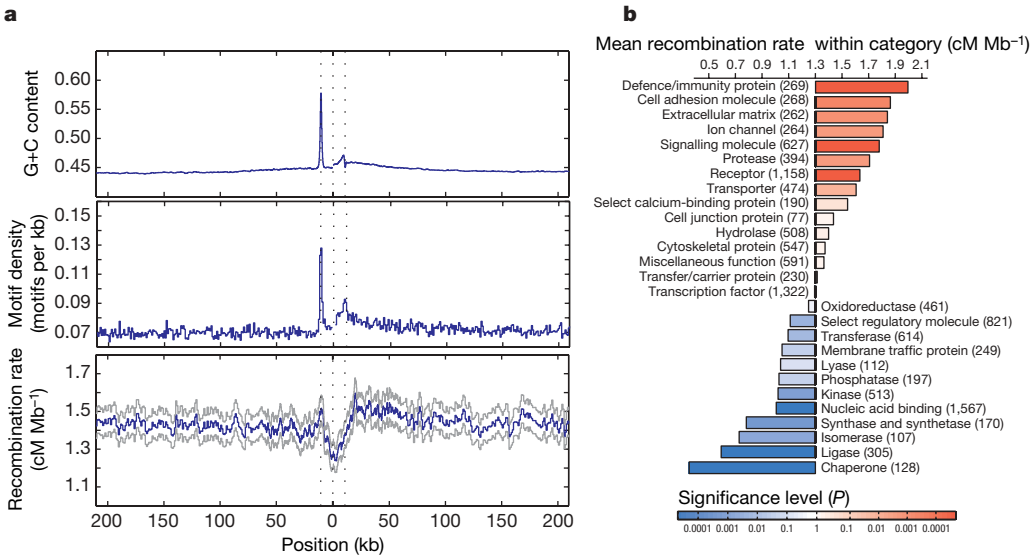


Figure 5 | Recombination rates around genes. a, The recombination rate, density of recombination-hotspot-associated motifs (all motifs with up to 1 bp different from the consensus CCTCCCTNNCCAC) and G+C content around genes. The blue line indicates the mean. For the recombination rate, grey lines indicate the quartiles of the distribution. Values were calculated separately 5′ from the transcription start site (the first dotted line) and 3′ from the transcription end site (third dotted line) and were joined at the median midpoint position of the transcription unit (central dotted line). Note the sharp drop in recombination rate within the transcription unit, the local increase around the transcription start site and the broad decrease away from the 3′ end of genes. These patterns only partly reflect the distribution of G+C content and the hotspot-associated motif, suggesting that additional factors influence recombination rates around genes. **b**, Recombination rates within genes of different molecular function[41]. The chart shows the increase or decrease for each category compared to the genome average. *P* values were estimated by permutation of category; numbers of genes are shown in parentheses.

and neutral SNPs, these results are consistent with a scenario in which the higher differentiation of non-synonymous SNPs is primarily driven by a reduction in the strength or efficacy of purifying selection in non-African populations.

Discussion and prospects

The International HapMap Project has been instrumental in making well-powered, large-scale, genome-wide association studies a reality. It is now clear that the HapMap can be a useful resource for the design and analysis of disease association studies in populations across the world[50–53]. Furthermore, the decreasing costs and increasing SNP density of standard genotyping panels mean that the focus of attention in disease association studies is shifting from candidate gene approaches towards genome-wide analyses. Alongside developments in technology, new statistical methodologies aimed at improving aspects of analysis, such as genotype calling[21,54], the identification of and correction for population stratification and relatedness[55,56], and imputation of untyped variants[21–23], are increasing the accuracy and reliability of genome-wide association studies.

Within this context, it is important to consider the future of the HapMap Project. Currently, additional samples from the populations used to develop the initial HapMap, as well as samples from seven additional populations (Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Tuscans in Italy; Gujarati Indian in Houston, Texas, USA; Denver (Colorado) metropolitan Chinese community; people of Mexican origin in Los Angeles, California, USA; and people with African ancestry in the southwestern United States; http://ccr.coriell.org/Sections/Collections/NHGRI/?SsId=11) will be sequenced and genotyped extensively to extend the HapMap, providing information on rarer variants and helping to enable genome-wide association studies in additional populations. There are also ongoing efforts by many groups to characterize additional forms of genetic variation, such as structural variation, and molecular phenotypes in the HapMap samples. Finally, in the future, whole-genome sequencing will provide a natural convergence of technologies to type both SNP and structural variation. Nevertheless, until that point, and even after, the HapMap Project data will provide an invaluable resource for understanding the structure of human genetic variation and its link to phenotype.

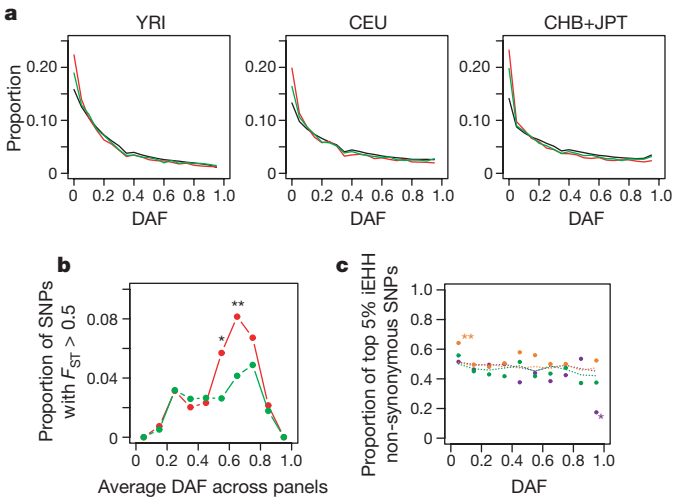


Figure 6 | Properties of non-synonymous and synonymous SNPs. a, The derived allele frequency (DAF) spectrum in each analysis panel for all SNPs (black), synonymous SNPs (green) and non-synonymous SNPs (red). Note the excess of rare variants for coding sequence SNPs but no excess of high-frequency derived variants. **b**, Enrichment of non-synonymous SNPs among genic SNPs showing high differentiation. For each of ten classes of derived allele frequency (averaged across analysis panels) the fraction of non-synonymous (red) and synonymous (green) variants in that class that show $F_{ST} > 0.5$ is shown. Note the strong enrichment of non-synonymous SNPs among SNPs of moderate to high derived-allele frequency (asterisk, $P < 0.05$; double asterisk, $P < 0.01$). **c**, Lack of enrichment of non-synonymous SNPs among those showing long-range haplotype structure. The integrated extended haplotype homozygosity (iEHH) statistic[9] was calculated for non-synonymous and synonymous SNPs in each analysis panel (YRI, green; CEU, orange; CHB+JPT, purple). For each of ten derived allele frequency classes, the proportion of non-synonymous SNPs among those showing the 5% most extreme statistics (within the allele frequency class) is shown (points). Also shown is the proportion of non-synonymous SNPs among SNPs in the coding sequence for each frequency class (dotted lines). Differences between synonymous and non-synonymous SNPs are tested for using a contingency table test.

METHODS SUMMARY

Of approximately 6.9 million SNPs in dbSNP release 122 approximately 4.7 million were selected for genotyping by Perlegen. 2.5 million SNPs were excluded because no assay could be designed and a further 350,000 were excluded for other reasons (see Methods). Perlegen performed genotyping using custom high-density oligonucleotide arrays as previously described[15]. Additional genotype submissions are described in the text. QC filters were applied as previously described[3]. Where multiple submissions met the QC criteria the submission with the lowest missing data rate was chosen for inclusion in the non-redundant filtered data set. Haplotypes were estimated from genotype data as described previously[3]. Ancestral states at SNPs were inferred by parsimony by comparison to orthologous bases in the chimpanzee (panTro2) and rhesus macaque (rheMac2) assemblies. Recombination rates and the location of recombination hotspots were estimated as described previously[3]. Additional details can be found in the Methods section and the Supplementary Information. The data described in this paper are in release 21 of the International HapMap Project.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 April; accepted 18 September 2007.

1. The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nature Rev. Genet.* **5,** 467–475 (2004).
2. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
3. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
4. Bowcock, A. M. Genomics: guilt by association. *Nature* **447,** 645–646 (2007).
5. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nature Genet.* **39,** 813–815 (2007).
6. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324 (2005).
7. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature Genet.* **38,** 86–92 (2006).
8. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genet.* **38,** 75–81 (2006).
9. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).
10. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
11. de Bakker, P. I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genet.* **38,** 1166–1172 (2006).
12. Pastinen, T. *et al.* Mapping common regulatory variants to human haplotypes. *Hum. Mol. Genet.* **14,** 3963–3971 (2005).
13. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1,** e78 (2005).
14. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437,** 1365–1369 (2005).
15. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307,** 1072–1079 (2005).
16. de Bakker, P. I. *et al.* Efficiency and power in genetic association studies. *Nature Genet.* **37,** 1217–1223 (2005).
17. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genet.* **38,** 663–667 (2006).
18. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* **38,** 659–662 (2006).
19. Burdick, J. T., Chen, W. M., Abecasis, G. R. & Cheung, V. G. *In silico* method for inferring genotypes in pedigrees. *Nature Genet.* **38,** 1002–1004 (2006).
20. Servin, B. R. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3,** e114 (2007).

21. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–668 (2007).
22. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316,** 1341–1345 (2007).
23. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genet.* **39,** 906–913 (2007).
24. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56,** 18–31 (2003).
25. Paabo, S. The mosaic that is our genome. *Nature* **421,** 409–412 (2003).
26. McVean, G., Spencer, C. C. & Chaix, R. Perspectives on human genetic variation from the HapMap Project. *PLoS Genet.* **1,** e54 (2005).
27. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
28. Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am. J. Hum. Genet.* **65,** 1493–1500 (1999).
29. Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15,** 789–795 (2006).
30. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236,** 1567–1570 (1987).
31. Leutenegger, A. L. *et al.* Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome. *Am. J. Hum. Genet.* **79,** 62–66 (2006).
32. Te Meerman, G. J., Van der Meulen, M. A. & Sandkuijl, L. A. Perspectives of identity by descent (IBD) mapping in founder populations. *Clin. Exp. Allergy* **25** (Suppl 2), 97–102 (1995).
33. Houwen, R. H. *et al.* Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* **8,** 380–386 (1994).
34. Durham, L. K. & Feingold, E. Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *Am. J. Hum. Genet.* **61,** 830–842 (1997).
35. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genet.* **36,** 151–156 (2004).
36. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304,** 581–584 (2004).
37. Myers, S. *et al.* The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.* **34,** 526–530 (2006).
38. Spencer, C. C. *et al.* The influence of recombination on human genetic diversity. *PLoS Genet.* **2,** e148 (2006).
39. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2,** 360–369 (2001).
40. Smith, A. V., Thomas, D. J., Munro, H. M. & Abecasis, G. R. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.* **15,** 1519–1534 (2005).
41. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13,** 2129–2141 (2003).
42. Winckler, W. *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308,** 107–111 (2005).
43. Ptak, S. E. *et al.* Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genet.* **37,** 429–434 (2005).
44. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
45. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* doi:10.1038/nature06250 (this issue).
46. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437,** 1153–1157 (2005).
47. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22,** 231–238 (1999).
48. Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12,** 1805–1814 (2002).
49. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312,** 1614–1620 (2006).
50. de Bakker, P. I. *et al.* Transferability of tag SNPs in genetic association studies in multiple populations. *Nature Genet.* **38,** 1298–1303 (2006).
51. Conrad, D. F. *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genet.* **38,** 1251–1260 (2006).
52. Service, S., Sabatti, C. & Freimer, N. Tag SNPs chosen from HapMap perform well in several population isolates. *Genet. Epidemiol.* **31,** 189–194 (2007).
53. Lim, J. *et al.* Comparative study of the linkage disequilibrium of an ENCODE region, chromosome 7p15, in Korean, Japanese, and Han Chinese samples. *Genomics* **87,** 392–398 (2006).
54. Rabbee, N. & Speed, T. P. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22,** 7–12 (2006).
55. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
56. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38,** 904–909 (2006).
57. Smith, R. A., Ho, P. J., Clegg, J. B., Kidd, J. R. & Thein, S. L. Recombination breakpoints in the human β-globin gene cluster. *Blood* **92,** 4415–4421 (1998).
58. Holloway, K., Lawson, V. E. & Jeffreys, A. J. Allelic recombination and *de novo* deletions in sperm in the human β-globin gene region. *Hum. Mol. Genet.* **15,** 1099–1111 (2006).
59. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38,** 1358–1370 (1984).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank many people who contributed to this project: all members of the genotyping laboratory and the sample, primer, bioinformatics, data quality and IT groups at Perlegen Sciences for technical and infrastructural support; J. Beck, C. Beiswanger, D. Coppock, A. Leach, J. Mintzer and L. Toji for transforming the Yoruba, Japanese and Han Chinese samples, distributing the DNA and cell lines, storing the samples for use in future research, and producing the community newsletters and reports; J. Greenberg and R. Anderson for providing funding and support for cell line transformation and storage in the NIGMS Human Genetic Cell Repository at the Coriell Institute; T. Dibling, T. Ishikura, S. Kanazawa, S. Mizusawa and S. Saito for help with genotyping; C. Hind and A. Moghadam for technical support in genotyping and all members of the subcloning and sequencing teams at the Wellcome Trust Sanger Institute; X. Ke for help with data analysis; Oxford E-Science Centre for provision of high-performance computing resources; H. Chen, W. Chen, L. Deng, Y. Dong, C. Fu, L. Gao, H. Geng, J. Geng, M. He, H. Li, H. Li, S. Li, X. Li, B. Liu, Z. Liu, F. Lu, F. Lu, G. Lu, C. Luo, X. Wang, Z. Wang, C. Ye and X. Yu for help with genotyping and sample collection; X. Feng, Y. Li, J. Ren and X. Zhou for help with sample collection; J. Fan, W. Gu, W. Guan, S. Hu, H. Jiang, R. Lei, Y. Lin, Z. Niu, B. Wang, L. Yang, W. Yang, Y. Wang, Z. Wang, S. Xu, W. Yan, H. Yang, W. Yuan, C. Zhang, J. Zhang, K. Zhang and G. Zhao for help with genotyping; P. Fong, C. Lai, C. Lau, T. Leung, L. Luk and W. Tong for help with genotyping; C. Pang for help with genotyping; K. Ding, B. Qiang, J. Zhang, X. Zhang and K. Zhou for help with genotyping; Q. Fu, S. Ghose, X. Lu, D. Nelson, A. Perez, S. Poole, R. Vega and H. Yonath for help with genotyping; C. Bruckner, T. Brundage, S. Chow, O. Iartchouk, M. Jain, M. Moorhead and K. Tran for help with genotyping; N. Addleman, J. Atilano, T. Chan, C. Chu, C. Ha, T. Nguyen, M. Minton and A. Phong for help with genotyping, and D. Lind for help with quality control and experimental design; R. Donaldson and S. Duan for help with genotyping, and J. Rice and N. Saccone for help with experimental design; J. Wigginton for help with implementing and testing QA/QC software; A. Clark, B. Keats, R. Myers, D. Nickerson and A. Williamson for providing advice to NIH; C. Juenger, C. Bennet, C. Bird, J. Melone, P. Nailer, M. Weiss, J. Witonsky and E. DeHaut-Combs for help with project management; M. Gray for organizing phone calls and meetings; D. Leja for help with figures; the Yoruba people of Ibadan, Nigeria, the people of Tokyo, Japan, and the community at Beijing Normal University, who participated in public consultations and community engagements; the people in these communities who donated their blood samples; and the people in the Utah CEPH community who allowed the samples they donated earlier to be used for the Project. This work was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology, the Wellcome Trust, Nuffield Trust, Wolfson Foundation, UK EPSRC, Genome Canada, Génome Québec, the Chinese Academy of Sciences, the Ministry of Science and Technology of the People's Republic of China, the National Natural Science Foundation of China, the Hong Kong Innovation and Technology Commission, the University Grants Committee of Hong Kong, the SNP Consortium, the US National Institutes of Health (FIC, NCI, NCRR, NEI, NHGRI, NIA, NIAAA, NIAID, NIAMS, NIBIB, NIDA, NIDCD, NIDCR, NIDDK, NIEHS, NIGMS, NIMH, NINDS, NLM, OD), the W.M. Keck Foundation, and the Delores Dore Eccles Foundation. All SNPs genotyped within the HapMap Project are available from dbSNP (http://www.ncbi.nlm.nih.gov/SNP); all genotype information is available from dbSNP and the HapMap website (http://www.hapmap.org).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.M. (mcvean@stats.ox.ac.uk) or M.D. (mjdaly@chgr.mgh.harvard.edu).

The International HapMap Consortium (Participants are arranged by institution and then alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.)

Genotyping centres: Perlegen Sciences Kelly A. Frazer (Principal Investigator)[1], Dennis G. Ballinger[2], David R. Cox[2], David A. Hinds[2], Laura L. Stuve[2]; **Baylor College of Medicine and ParAllele BioScience** Richard A. Gibbs (Principal Investigator)[3], John W. Belmont[3], Andrew Boudreau[4], Paul Hardenbol[5], Suzanne M. Leal[3], Shiran Pasternak[6], David A. Wheeler[3], Thomas D. Willis[4], Fuli Yu[7]; **Beijing Genomics Institute** Huanming Yang (Principal Investigator)[8], Changqing Zeng (Principal Investigator)[8], Yang Gao[8], Haoran Hu[8], Weitao Hu[8], Chaohua Li[8], Wei Lin[8], Siqi Liu[8], Hao Pan[8], Xiaoli Tang[8], Jian Wang[8], Wei Wang[8], Jun Yu[8], Bo Zhang[8], Qingrun Zhang[8], Hongbin Zhao[8], Hui Zhao[8], Jun Zhou[8]; **Broad Institute of Harvard and Massachusetts Institute of Technology**

Stacey B. Gabriel (Project Leader)[7], Rachel Barry[7], Brendan Blumenstiel[7], Amy Camargo[7], Matthew Defelice[7], Maura Faggart[7], Mary Goyette[7], Supriya Gupta[7], Jamie Moore[7], Huy Nguyen[7], Robert C. Onofrio[7], Melissa Parkin[7], Jessica Roy[7], Erich Stahl[7], Ellen Winchester[7], Liuda Ziaugra[7], David Altshuler (Principal Investigator)[7,9]; **Chinese National Human Genome Center at Beijing** Yan Shen (Principal Investigator)[10], Zhijian Yao[10]; **Chinese National Human Genome Center at Shanghai** Wei Huang (Principal Investigator)[11], Xun Chu[11], Yungang He[11], Li Jin[12], Yangfan Liu[11], Yayun Shen[11], Weiwei Sun[11], Haifeng Wang[11], Yi Wang[11], Ying Wang[11], Xiaoyan Xiong[11], Liang Xu[11]; **Chinese University of Hong Kong** Mary M. Y. Waye (Principal Investigator)[13], Stephen K. W. Tsui[13]; **Hong Kong University of Science and Technology** Hong Xue (Principal Investigator)[14], J. Tze-Fei Wong[14]; **Illumina** Luana M. Galver (Project Leader)[15], Jian-Bing Fan[15], Kevin Gunderson[15], Sarah S. Murray[1], Arnold R. Oliphant[16], Mark S. Chee (Principal Investigator)[17]; **McGill University and Génome Québec Innovation Centre** Alexandre Montpetit (Project Leader)[18], Fanny Chagnon[18], Vincent Ferretti[18], Martin Leboeuf[18], Jean-François Olivier[4], Michael S. Phillips[18], Stéphanie Roumy[15], Clémentine Sallée[19], Andrei Verner[18], Thomas J. Hudson (Principal Investigator)[20]; **University of California at San Francisco and Washington University** Pui-Yan Kwok (Principal Investigator)[21], Dongmei Cai[21], Daniel C. Koboldt[22], Raymond D. Miller[22], Ludmila Pawlikowska[21], Patricia Taillon-Miller[22], Ming Xiao[21]; **University of Hong Kong** Lap-Chee Tsui (Principal Investigator)[23], William Mak[23], You Qiang Song[23], Paul K. H. Tam[23]; **University of Tokyo and RIKEN** Yusuke Nakamura (Principal Investigator)[24,25], Takahisa Kawaguchi[25], Takuya Kitamoto[25], Takashi Morizono[25], Atsushi Nagashima[25], Yozo Ohnishi[25], Akihiro Sekine[25], Toshihiro Tanaka[25], Tatsuhiko Tsunoda[25]; **Wellcome Trust Sanger Institute** Panos Deloukas (Project Leader)[26], Christine P. Bird[26], Marcos Delgado[26], Emmanouil T. Dermitzakis[26], Rhian Gwilliam[26], Sarah Hunt[26], Jonathan Morrison[27], Don Powell[26], Barbara E. Stranger[26], Pamela Whittaker[26], David R. Bentley (Principal Investigator)[28]

Analysis groups: Broad Institute Mark J. Daly (Project Leader)[7,9], Paul I. W. de Bakker[7,9], Jeff Barrett[7,9], Yves R. Chretien[7], Julian Maller[7,9], Steve McCarroll[7,9], Nick Patterson[7], Itsik Pe'er[29], Alkes Price[7], Shaun Purcell[9], Daniel J. Richter[7], Pardis Sabeti[7], Richa Saxena[7,9], Stephen F. Schaffner[7], Pak C. Sham[23], Patrick Varilly[7], David Altshuler (Principal Investigator)[7,9]; **Cold Spring Harbor Laboratory** Lincoln D. Stein (Principal Investigator)[6], Lalitha Krishnan[6], Albert Vernon Smith[6], Marcela K. Tello-Ruiz[6], Gudmundur A. Thorisson[30]; **Johns Hopkins University School of Medicine** Aravinda Chakravarti (Principal Investigator)[31], Peter E. Chen[31], David J. Cutler[31], Carl S. Kashuk[31], Shin Lin[31]; **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[32], Weihua Guan[32], Yun Li[32], Heather M. Munro[33], Zhaohui Steve Qin[32], Daryl J. Thomas[34]; **University of Oxford** Gilean McVean (Project Leader)[35], Adam Auton[35], Leonardo Bottolo[35], Niall Cardin[35], Susana Eyheramendy[35], Colin Freeman[35], Jonathan Marchini[35], Simon Myers[35], Chris Spencer[7], Matthew Stephens[36], Peter Donnelly (Principal Investigator)[35]; **University of Oxford, Wellcome Trust Centre for Human Genetics** Lon R. Cardon (Principal Investigator)[37], Geraldine Clarke[38], David M. Evans[38], Andrew P. Morris[38], Bruce S. Weir[39]; **RIKEN** Tatsuhiko Tsunoda (Principal Investigator)[25], Todd A. Johnson[25]; **US National Institutes of Health** James C. Mullikin[40]; **US National Institutes of Health National Center for Biotechnology Information** Stephen T. Sherry[41], Michael Feolo[41], Andrew Skol[42]

Community engagement/public consultation and sample collection groups: Beijing Normal University and Beijing Genomics Institute Houcan Zhang[43], Changqing Zeng[8], Hui Zhao[8]; **Health Sciences University of Hokkaido, Eubios Ethics Institute, and Shinshu University** Ichiro Matsuda (Principal Investigator)[44], Yoshimitsu Fukushima[45], Darryl R. Macer[46], Eiko Suda[47]; **Howard University and University of Ibadan** Charles N. Rotimi (Principal Investigator)[48], Clement A. Adebamowo[49], Ike Ajayi[49], Toyin Aniagwu[49], Patricia A. Marshall[50], Chibuzor Nkwodimmah[49], Charmaine D. M. Royal[48]; **University of Utah** Mark F. Leppert (Principal Investigator)[51], Missy Dixon[51], Andy Peiffer[51]

Ethical, legal and social issues: Chinese Academy of Social Sciences Renzong Qiu[52]; **Genetic Interest Group** Alastair Kent[53]; **Kyoto University** Kazuto Kato[54]; **Nagasaki University** Norio Niikawa[55]; **University of Ibadan School of Medicine** Isaac F. Adewole[49]; **University of Montréal** Bartha M. Knoppers[19]; **University of Oklahoma** Morris W. Foster[56]; **Vanderbilt University** Ellen Wright Clayton[57]; **Wellcome Trust** Jessica Watkin[58]

SNP discovery: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[3], John W. Belmont[3], Donna Muzny[3], Lynne Nazareth[3], Erica Sodergren[3], George M. Weinstock[3], David A. Wheeler[3], Imtaz Yakub[3]; **Broad Institute of Harvard and Massachusetts Institute of Technology** Stacey B. Gabriel (Project Leader)[7], Robert C. Onofrio[7], Daniel J. Richter[7], Liuda Ziaugra[7], Bruce W. Birren[7], Mark J. Daly[7,9], David Altshuler (Principal Investigator)[7,9]; **Washington University** Richard K. Wilson (Principal Investigator)[59], Lucinda L. Fulton[59]; **Wellcome Trust Sanger Institute** Jane Rogers (Principal Investigator)[26], John Burton[26], Nigel P. Carter[26], Christopher M. Clee[26], Mark Griffiths[26], Matthew C. Jones[26], Kirsten McLay[26], Robert W. Plumb[26], Mark T. Ross[26], Sarah K. Sims[26], David L. Willey[26]

Scientific management: Chinese Academy of Sciences Zhu Chen[60], Hua Han[60], Le Kang[60]; **Genome Canada** Martin Godbout[61], John C. Wallenburg[62]; **Génome Québec** Paul L'Archevêque[63], Guy Bellemare[63]; **Japanese Ministry of Education, Culture, Sports, Science and Technology** Koji Saeki[64]; **Ministry of Science and Technology of the People's Republic of China** Hongguang Wang[65], Daochang An[65], Hongbo Fu[65], Qing Li[65], Zhen Wang[65]; **The Human Genetic Resource Administration of China** Renwu Wang[66]; **The SNP Consortium** Arthur L. Holden[15]; **US National Institutes of Health** Lisa D. Brooks[67], Jean E. McEwen[67], Mark S. Guyer[67], Vivian Ota Wang[67,68], Jane L. Peterson[67], Michael Shi[69], Jack Spiegel[70], Lawrence M. Sung[71], Lynn F. Zacharia[67], Francis S. Collins[72]; **Wellcome Trust** Karen Kennedy[61], Ruth Jamieson[58], John Stewart[58]

[1]The Scripps Research Institute, 10550 North Torrey Pines Road MEM275, La Jolla, California 92037, USA. [2]Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, California 94043, USA. [3]Baylor College of Medicine, Human Genome Sequencing Center, Department of Molecular and Human Genetics, 1 Baylor Plaza, Houston, Texas 77030, USA. [4]Affymetrix, Inc., 3420 Central Expressway, Santa Clara, California 95051, USA. [5]Pacific Biosciences, 1505 Adams Drive, Menlo Park, California 94025, USA. [6]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [7]The Broad Institute of Harvard and Massachusetts Institute of Technology, 1 Kendall Square, Cambridge, Massachusetts 02139, USA. [8]Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 100300, China. [9]Massachusetts General Hospital and Harvard Medical School, Simches Research Center, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [10]Chinese National Human Genome Center at Beijing, 3-707 N. Yongchang Road, Beijing Economic-Technological Development Area, Beijing 100176, China. [11]Chinese National Human Genome Center at Shanghai, 250 Bi Bo Road, Shanghai 201203, China. [12]Fudan University and CAS-MPG Partner Institute for Computational Biology, School of Life Sciences, SIBS, CAS, Shanghai 201203, China. [13]The Chinese University of Hong Kong, Department of Biochemistry, The Croucher Laboratory for Human Genetics, 6/F Mong Man Wai Building, Shatin, Hong Kong. [14]Hong Kong University of Science and Technology, Department of Biochemistry and Applied Genomics Center, Clear Water Bay, Knowloon, Hong Kong. [15]Illumina, 9885 Towne Centre Drive, San Diego, California 92121, USA. [16]Complete Genomics, Inc., 658 North Pastoria Avenue, Sunnyvale, California 94085, USA. [17]Prognosys Biosciences, Inc., 4215 Sorrento Valley Boulevard, Suite 105, San Diego, California 92121, USA. [18]McGill University and Génome Québec Innovation Centre, 740 Dr. Penfield Avenue, Montréal, Québec H3A 1A4, Canada. [19]University of Montréal, The Public Law Research Centre (CRDP), PO Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada. [20]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 500, Toronto, Ontario M5G 1L7, Canada. [21]University of California, San Francisco, Cardiovascular Research Institute, 513 Parnassus Avenue, Box 0793, San Francisco, California 94143, USA. [22]Washington University School of Medicine, Department of Genetics, 660 South Euclid Avenue, Box 8232, St Louis, Missouri 63110, USA. [23]University of Hong Kong, Genome Research Centre, 6/F, Laboratory Block, 21 Sassoon Road, Pokfulam, Hong Kong. [24]University of Tokyo, Institute of Medical Science, 4-6-1 Sirokanedai, Minato-ku, Tokyo 108-8639, Japan. [25]RIKEN SNP Research Center, 1-7-22 Suehiro-cho, Tsurumi-ku Yokohama, Kanagawa 230-0045, Japan. [26]Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [27]University of Cambridge, Department of Oncology, Cambridge CB1 8RN, UK. [28]Solexa Ltd, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK. [29]Columbia University, 500 West 120th Street, New York, New York 10027, USA. [30]University of Leicester, Department of Genetics, Leicester LE1 7RH, UK. [31]Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Broadway Research Building, Suite 579, 733 North Broadway, Baltimore, Maryland 21205, USA. [32]University of Michigan, Center for Statistical Genetics, Department of Biostatistics, 1420 Washington Heights, Ann Arbor, Michigan 48109, USA. [33]International Epidemiology Institute, 1455 Research Boulevard, Suite 550, Rockville, Maryland 20850, USA. [34]Center for Biomolecular Science and Engineering, Engineering 2, Suite 501, Mail Stop CBSE/ITI, UC Santa Cruz, Santa Cruz, California 95064, USA. [35]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK. [36]University of Chicago, Department of Statistics, 5734 South University Avenue, Eckhart Hall, Room 126, Chicago, Illinois 60637, USA. [37]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. [38]University of Oxford/Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [39]University of Washington Department of Biostatistics, Box 357232, Seattle, Washington 98195, USA. [40]US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA. [41]US National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. [42]University of Chicago, Department of Medicine, Section of Genetic Medicine, 5801 South Ellis, Chicago, Illinois 60637, USA. [43]Beijing Normal University, 19 Xinjiekouwai Street, Beijing 100875, China. [44]Health Sciences University of Hokkaido, Ishikari Tobetsu Machi 1757, Hokkaido 061-0293, Japan. [45]Shinshu University School of Medicine, Department of Medical Genetics, Matsumoto 390-8621, Japan. [46]United Nations Educational, Scientific and Cultural Organization (UNESCO Bangkok), 920 Sukhumwit Road, Prakanong, Bangkok 10110, Thailand. [47]University of Tsukuba, Eubios Ethics Institute, PO Box 125, Tsukuba Science City 305-8691, Japan. [48]Howard University, National Human Genome Center, 2216 6th Street, NW, Washington DC 20059, USA. [49]University of Ibadan College of Medicine, Ibadan, Oyo State, Nigeria. [50]Case Western Reserve University School of Medicine, Department of Bioethics, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. [51]University of Utah, Eccles Institute of Human Genetics, Department of Human Genetics, 15 North 2030 East, Salt Lake City, Utah 84112, USA. [52]Chinese Academy of Social Sciences, Institute of Philosophy/Center for Applied Ethics, 2121, Building 9, Caoqiao Xinyuan 3 Qu, Beijing 100067, China. [53]Genetic Interest Group, 4D Leroy House, 436 Essex Road, London N130P, UK. [54]Kyoto University, Institute for Research in Humanities and Graduate School of Biostudies, Ushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan. [55]Nagasaki University Graduate

School of Biomedical Sciences, Department of Human Genetics, Sakamoto 1-12-4, Nagasaki 852-8523, Japan. [56]University of Oklahoma, Department of Anthropology, 455 West Lindsey Street, Norman, Oklahoma 73019, USA. [57]Vanderbilt University, Center for Genetics and Health Policy, 507 Light Hall, Nashville, Tennessee 37232, USA. [58]Wellcome Trust, 215 Euston Road, London NW1 2BE, UK. [59]Washington University School of Medicine, Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [60]Chinese Academy of Sciences, 52 Sanlihe Road, Beijing 100864, China. [61]Genome Canada, 150 Metcalfe Street, Suite 2100, Ottawa, Ontario K2P 1P1, Canada. [62]McGill University, Office of Technology Transfer, 3550 University Street, Montréal, Québec H3A 2A7, Canada. [63]Génome Québec, 630, boulevard René-Lévesque Ouest, Montréal, Québec H3B 1S6, Canada. [64]Ministry of Education, Culture, Sports, Science, and Technology, 3-2-2 Kasumigaseki, Chiyodaku, Tokyo 100-8959, Japan. [65]Ministry of Science and Technology of the People's Republic of China, 15 B. Fuxing Road, Beijing 100862, China. [66]The Human Genetic Resource Administration of China, b7, Zaojunmiao, Haidian District, Beijing 100081, China. [67]US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA. [68]US National Institutes of Health, Office of Behavioral and Social Science Research, 31 Center Drive, Bethesda, Maryland 20892, USA. [69]Novartis Pharmaceuticals Corporation, Biomarker Development, One Health Plaza, East Hanover, New Jersey 07936, USA. [70]US National Institutes of Health, Office of Technology Transfer, 6011 Executive Boulevard, Rockville, Maryland 20852, USA. [71]University of Maryland School of Law, 500 West Baltimore Street, Baltimore, Maryland 21201, USA. [72]US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.

METHODS

SNP selection and genotyping. All SNPs in dbSNP release 122 were considered for genotyping by Perlegen. Among these the following were excluded: SNPs for which no assay could be designed (primarily through location in repeat-rich regions; approximately 2.5 million); SNPs shown previously in samples from related populations[15] to be most probably in perfect association ($r^2 = 1$) with a Phase I SNP (approximately 122,000); all but one of SNPs shown previously[15] to be most probably in perfect association ($r^2 = 1$) with each other but not with a Phase I SNP (approximately 62,000); and SNPs shown previously[15] to have MAF < 0.05 (approximately 119,000). In addition, a few SNPs were excluded for efficiency (for example, if an amplicon contained a single SNP). Approximately 30,000 SNPs that had been typed in Phase I were deliberately retyped in Phase II to allow detailed comparisons of data quality, and an additional 15,000 SNPs that showed discrepancies between multiple genotyping attempts in Phase I were re-typed in Phase II. A further 2,000 SNPs identified by the Mammalian Gene Collection were also typed.

Perlegen performed genotyping using custom high-density oligonucleotide arrays as previously described[15]. Initially, a pilot phase was carried out on chromosome 2p to optimize experimental workflow and data handling. Details of amplicons used in the experiment and PCR primers can be found at http://genome.perlegen.com/pcr/ and also on the HapMap website. The arrays were tiled with sets of 25-bp probes for each SNP, with either 40 or 24 probes per SNP. These consisted of four sets of features, corresponding to forward and reverse strand tilings of sequences complementary to each of the two SNP alleles. Within a feature set, the position of the SNP within the oligonucleotide varied from position 11 to position 15. Mismatch probes were used to measure background, and by comparison with the perfect match probes, to detect the presence or absence of a specific PCR product. The 40-feature and 24-feature tilings both provided 10 perfect-match features for each SNP allele and differed only in the number of mismatch probes.

Genotypes were scored by clustering intensity measurements as previously described[15]. In addition, quality scores similar to Phred scores were computed for each genotype call, based on a combination of experimental metrics correlated to data quality. Assays with overall call rates less than 80% or with poor average quality scores were flagged as failed. About 38% of the tiled assays failed these basic criteria, and the remainder were processed using the more rigorous HapMap Project data quality control filters. For analysis of the whole genome, probes for 4,373,926 distinct SNPs were tiled onto 32 chip designs, with 32 SNPs tiled in replicate onto each chip design for quality control (QC). Perlegen did not type the samples by plates as had been done for the Phase I genotyping, instead typing large numbers of SNPs one sample at a time. Consequently, blank wells on each plate were not included as a component of QC for this genotyping. In the Phase I HapMap a single JPT sample had been excluded because of technical problems. Perlegen typed a replacement sample (from the original JPT collection) for all new SNPs. This sample was not specifically genotyped on the Phase I SNPs, although a substantial fraction of these was typed in Phase II.

Additional genotype submissions came from the Affymetrix GeneChip Human Mapping 500K array called with the BRLMM algorithm. In release 21a additional genotype submissions were incorporated from the MHC haplotype consortium[11], the Illumina HumanHap300 BeadChip, the Illumina Human-1 Genotyping BeadChip and the 10K non-synonymous SNP set from Affymetrix (ParAllele).

Details of primer design, DNA amplification, DNA labelling and hybridization and signal detection for the Perlegen platform can be found in Supplementary Text 7.

QC analyses. Genotype submissions were assessed for mendelian errors (where possible), missing data rates and Hardy–Weinberg proportions. QC filters were applied as previously described[3]; to achieve QC+ status a SNP had to have fewer than two mendelian errors, less than 20% missing data and $P > 0.001$ for Hardy–Weinberg analysis. The consensus data set consists only of SNPs for which QC+ submissions were available from all analysis panels. Where multiple submissions met the QC criteria the submission with the lowest missing data rate was chosen for inclusion in the non-redundant filtered data set. Comparison of the Phase II HapMap with the Affymetrix 500K genotypes has shown approximately 20 SNPs where the reported minor allele is discrepant (referred to as 'allele-flipping'). Over the entire data set, we expect that 500–2,000 SNPs have this problem and the vast majority will occur in SNPs from Phase I of the project. The Data Coordination Center (DCC) is working to resolve as many of these as possible.

Analyses of data quality. See Supplementary Text 2.

Analyses of population stratification, relatedness and homozygosity. See Supplementary Texts 3–6.

Analysis of recombination rate and gene ontology. We used the Panther Database[41] to obtain details of the gene molecular function and biological process. Genes are grouped into 28 top-level molecular function groups and 30 top-level biological process groups, with each gene allowed to exist in more than one group. We identified 14,979 non-overlapping autosomal genes from the Panther RefSeq Annotation for which we could obtain recombination rates. Of these, 9,735 had at least one assigned molecular function and 9,432 had at least one assigned biological process. Genes without a molecular function or biological process were removed from the corresponding analysis. To control for gene size, we estimated the mean recombination rate over a 20-kb region centred on the mid-point of each gene transcription region.

Genes were grouped based on molecular function and biological process. A mean recombination rate was calculated for each group. The significance of the result from each group was calculated via a permutation test involving 10^5 random groupings of genes. No correction was made for multiple testing. To account for the effect of G+C content on recombination, we performed a linear regression between the G+C content and recombination rate of all genes in each sample. Using the estimated regression parameters, the proportion of recombination explained by G+C content was subtracted from each gene.

Identification of non-synonymous SNPs and tests for natural selection. Using annotations from dbSNP release 125 we identified 17,427 polymorphic non-synonymous SNPs in release 21 and 15,976 polymorphic synonymous SNPs. Of these, 15,583 non-synonymous and 14,324 synonymous SNPs were autosomal and could have ancestral allele status unambiguously assigned by parsimony through comparison to the chimpanzee and macaque genomes. We used the phased haplotypes for analysis in which missing data had been imputed. F_{ST} was calculated using the method of Weir and Cockerham[59].

To detect recent partial selective sweeps we used the long-range haplotype (LRH) test[44,49] and the integrated haplotype score (iHS) test[9]. On simulated data[45], we found that the tests have similar power to detect recent selection but the iHS test has slightly lower power at low haplotype frequency and the LRH test has slightly lower power at high frequency. This can be seen in applications to HapMap Phase I data[3,9], where the iHS test misses the well-known cases of *HBB* and *CD36* and the LRH test misses the *SULT1C2* region. Although both tests are based on the concept of EHH[44], we observed that the false positives produced by the two tests tend not to overlap and thus that signals detected by both tests have a very low false-positive rate.

ARTICLES

Helicobacter exploits integrin for type IV secretion and kinase activation

Terry Kwok[1]†, Dana Zabler[1], Sylwia Urman[3], Manfred Rohde[4], Roland Hartig[2], Silja Wessler[5], Rolf Misselwitz[6]†, Jürgen Berger[7], Norbert Sewald[3], Wolfgang König[1] & Steffen Backert[1]

Integrins are important mammalian receptors involved in normal cellular functions as well as pathogenesis of chronic inflammation and cancer. We propose that integrins are exploited by the gastric pathogen and type-1 carcinogen *Helicobacter pylori* for injection of the bacterial oncoprotein cytotoxin-associated gene A (CagA) into gastric epithelial cells. Virulent *H. pylori* express a type-IV secretion pilus that injects CagA into the host cell; CagA then becomes tyrosine-phosphorylated by Src family kinases. However, the identity of the host cell receptor involved in this process has remained unknown. Here we show that the *H. pylori* CagL protein is a specialized adhesin that is targeted to the pilus surface, where it binds to and activates integrin $\alpha_5\beta_1$ receptor on gastric epithelial cells through an arginine-glycine-aspartate motif. This interaction triggers CagA delivery into target cells as well as activation of focal adhesion kinase and Src. Our findings provide insights into the role of integrins in *H.-pylori*-induced pathogenesis. CagL may be exploited as a new molecular tool for our further understanding of integrin signalling.

Integrins are cell adhesion receptors that mediate cell–cell, cell–extracellular matrix and cell–pathogen interactions[1–7]. Integrins exist as heterodimers of α- and β-subunits; in mammals there are 18 α-subunits and 8 β-subunits identified that form 24 distinct heterodimers[1–4]. Integrin function is dynamically regulated by processes termed outside-in and inside-out signalling. Integrin-mediated cellular adhesion occurs at sites known as focal adhesions, where integrin signalling is mediated predominantly by the protein tyrosine kinases FAK (focal adhesion kinase) and Src[1–4]. Deregulation of integrin functions can be associated with diseases including chronic inflammation, heart failure and cancer[1–4]. Furthermore, a number of bacteria and viruses have been found to bind integrin receptors for adhesion to or invasion into host cells[5–7]. Here we report that the gastric pathogen *H. pylori* exploits integrin for injection of virulence factors into mammalian cells and simultaneous tyrosine kinase activation.

A wide range of Gram-negative bacterial pathogens, including *H. pylori*, translocate virulence factors into host target cells by multi-subunit transport apparatuses known as type-IV secretion systems[8,9]. These type-IV secretion systems, which typically consist of a membrane-spanning secretion channel and an extracellular pilus, act as molecular syringes. The type-IV secretion system of *H. pylori* has attracted considerable interest because it is associated with the development of severe inflammation, peptic ulcer disease and gastric cancer in humans[9–13]. The *H. pylori* type-IV secretion, encoded by a 40-kilobase *cag* pathogenicity island (*cag*PAI), injects the CagA oncoprotein as well as peptidoglycan into host cells, resulting in activation of NF-κB and induction of potent pro-inflammatory chemokines such as interleukin (IL)-8 (refs 9–13). Translocated CagA undergoes tyrosine phosphorylation by Src, leading to actin-cytoskeletal rearrangements, scattering and elongation of infected host cells in cell culture[9–13]. These phenotypic changes resemble those of malignant cellular transformation and have been the subject of intensive studies[9–13]. The identity of the host cell receptor of the type-IV secretion system or the mechanism by which Src is activated during *H. pylori* infection are, however, unknown, hampering our understanding of the molecular mechanism by which *H. pylori* causes gastric ulcer and cancer.

Integrin $\alpha_5\beta_1$ is involved in type-IV secretion of CagA

To understand the molecular basis of CagA injection, we first investigated the subcellular localization of phosphorylated CagA (CagA-pY) shortly after attachment of *H. pylori* to AGS gastric epithelial cells using an antibody that recognizes CagA phosphorylated at the major phosphorylation site, tyrosine residue Y972 (ref. 14, Fig. 1 and Supplementary Fig. 1). CagA-pY was found to co-localize with FAK or vinculin at the focal adhesions (Fig. 1a, arrows, and data not shown), suggesting that CagA is either rapidly recruited to or injected across the host plasma membrane at the sites of focal adhesion. This led us to examine whether integrins, which are present in focal adhesions[1–4], might be the host receptors involved in the injection of CagA. The first line of evidence came from the observation that the *Escherichia coli* strain HB101 that expresses the integrin-β_1-binding *Yersinia* invasin InvA (HB101 *invA*$^+$)[6] inhibited CagA phosphorylation in AGS cells (Supplementary Fig. 2a), whereas wild-type HB101 had no effect on CagA phosphorylation (data not shown). Thus, InvA might be able to block injection of CagA by competing with the type-IV secretion apparatus of *H. pylori* for binding to integrin β_1 and in particular the β_1-chain integrin member $\alpha_5\beta_1$, which is abundantly expressed in AGS cells (Supplementary Fig. 3). In accordance with this hypothesis, the integrin β_1 and α_5 function-blocking antibodies AIIB2 and BIIG2, respectively, inhibited both *H.-pylori*-induced CagA phosphorylation and AGS cell scattering and elongation (Supplementary Fig. 2b, c), indicating that integrin β_1 and α_5 are

[1]Department of Medical Microbiology, and [2]Department of Immunology, Otto von Guericke University, Leipziger Strasse 44, D-39120 Magdeburg, Germany. [3]Department of Chemistry, Organic and Bioorganic Chemistry, Bielefeld University, Universitätsstrasse 25, D-33615 Bielefeld, Germany. [4]Helmholtz Center for Infection Research, Department of Microbial Pathogenesis, Inhoffen Strasse 7, D-38124 Braunschweig, Germany. [5]Paul Ehrlich Institute, Paul-Ehrlich-Strasse 51-59, D-63225 Langen, Germany. [6]Max Delbrück Center for Molecular Medicine, Robert-Roessle-Strasse 10, D-13125 Berlin, Germany. [7]Max Planck Institute for Developmental Biology, Spemannstrasse 35, D-72076 Tübingen, Germany. †Present addresses: University of Zürich, Institute of Medical Virology, Gloriastrasse 30/32, CH-8006 Zürich, Switzerland (T.K.); Institute for Immuno Genetics, Charité University Clinics Berlin, Humboldt University Berlin, Spandauer Damm 130, D-14050 Berlin, Germany (R.M.).

required for injection of CagA by *H. pylori*. Furthermore, no phosphorylation of CagA was seen in *H.-pylori*-infected integrin-β_1 knockout mouse fibroblasts (GD25)[15], whereas CagA phosphorylation was restored in GD25 fibroblasts in which integrin β_1 was re-expressed (GD25β1A) (Fig. 1b). The presence of CagA and CagA-pY in the cytosol of infected GD25β1A cells but not GD25 cells (Fig. 1b) further confirms that integrin $\alpha_5\beta_1$ is required for the type-IV secretion and phosphorylation of CagA during *H. pylori* infection of host target cells.

CagL is expressed on type-IV secretion pili

Many integrin ligands such as fibronectin and vitronectin carry the arginine-glycine-aspartate (RGD) motif, which serves as a recognition site for integrins[1–4]. Sequence analysis indicated that CagL, a protein highly conserved among pathogenic *H. pylori* strains, is the only *cag*PAI-encoded gene product that contains an RGD motif (Fig. 2a). Deletion of *cagL* significantly diminished the ability of *H. pylori* to induce secretion of IL-8 by host cells[12], indicating that CagL is essential for pathogenesis. We postulated that CagL might function as an interacting partner of integrin β_1 during type-IV secretion. Immunofluorescence microscopy and scanning electron microscopy (FESEM) showed that CagL is indeed expressed on the surface of *H. pylori* (Fig. 2b, c). Notably, within only 30 min of contact of *H. pylori* with AGS cells, numerous needle-like pili (100–200 nm in length) appeared on the bacterium in a CagL- and RGD-dependent manner (Fig. 2d–f and Supplementary Fig. 4a–c). CagL was found abundantly on these pili. Representative examples from 60 randomly selected pili are shown (Fig. 2d and Supplementary Fig. 5a); they are reminiscent of the agrobacterial and *H. pylori* type-IV secretion apparatus[8,9,16]. These structures were clearly not flagella, because pilus formation and CagL surface exposure were also observed in infections with the flagella-deficient mutant *H. pylori* Δ*flaA* (data not shown). In contrast, neither pilus formation nor CagL surface expression was seen in infections with Δ*virB7*, Δ*virB10* or Δ*virB11*, *H. pylori* mutants that are defective in type-IV secretion (data not shown). Results of immunogold labelling showed the presence of CagA signals at the tips of these pili, supporting the theory that CagA is delivered through these surface appendages (Fig. 2e and Supplementary Fig. 5b). Taken together, these findings led us to propose that CagL may function in an RGD-dependent manner as a sensor or adhesin on the *H. pylori* type-IV secretion pilus surface to trigger the injection of CagA on contact with host cells.

Functional role of CagL and its RGD motif

To investigate whether the RGD motif in CagL has a functional role in the injection of CagA, we infected AGS gastric epithelial cells with wild-type *H. pylori* and *cagL* mutants. A single amino acid substitution in the RGD motif, where either the aspartate or glycine residue is substituted by an alanine residue (CagL(RGA) or CagL(RAD), respectively), was sufficient to abolish CagA translocation into the cytosol, tyrosine phosphorylation of CagA, as well as the elongation phenotype (Fig. 3a–d). Complementation of the *H. pylori* Δ*cagL* deletion mutant with wild-type *cagL*, but not *cagL*(RGA) or *cagL*(RAD) mutants, restored the various phenotypes (Fig. 3a–d). These observations suggest that the RGD motif of CagL has an essential role in translocation and tyrosine phosphorylation of CagA. Furthermore, the CagL-derived RGD peptide c-(-Arg-Gly-Asp-D-Leu-Ala-), but not the control RAD peptide c-(-Arg-Ala-Asp-D-Leu-Ala-), complemented the ability of the *H. pylori cagL*(RAD) mutant by approximately 40–50% in triggering phosphorylation of CagA and phenotypic responses (Fig. 3d), as well as pili formation (Supplementary Fig. 4d). These findings strongly suggest that the RGD motif is essential for the concomitant activation of both

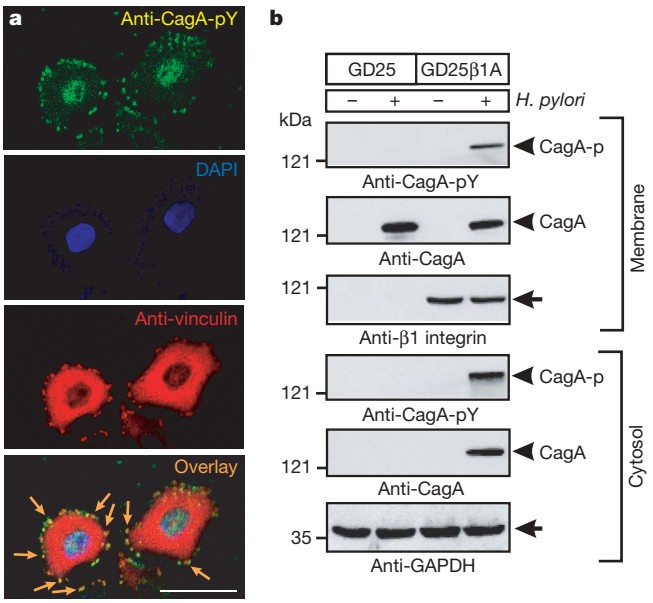


Figure 1 | Phosphorylated CagA co-localizes with focal adhesion proteins during infection with *H. pylori*, and integrin β_1 is required for type-IV secretion of CagA into target cells. **a**, Confocal microscopy of AGS cells infected with wild-type *H. pylori* for 2 h. Arrows indicate phosphorylated CagA at focal adhesions. Scale bar, 10 μm. **b**, Immunoblots of the membrane and cytosolic fractions of GD25 (integrin β_1 knockout) and GD25β1A (integrin β_1-expressing) mouse fibroblasts infected with *H. pylori* or PBS control. Bands corresponding to CagA are indicated by arrowheads and those of integrin β_1 and GAPDH by arrows.

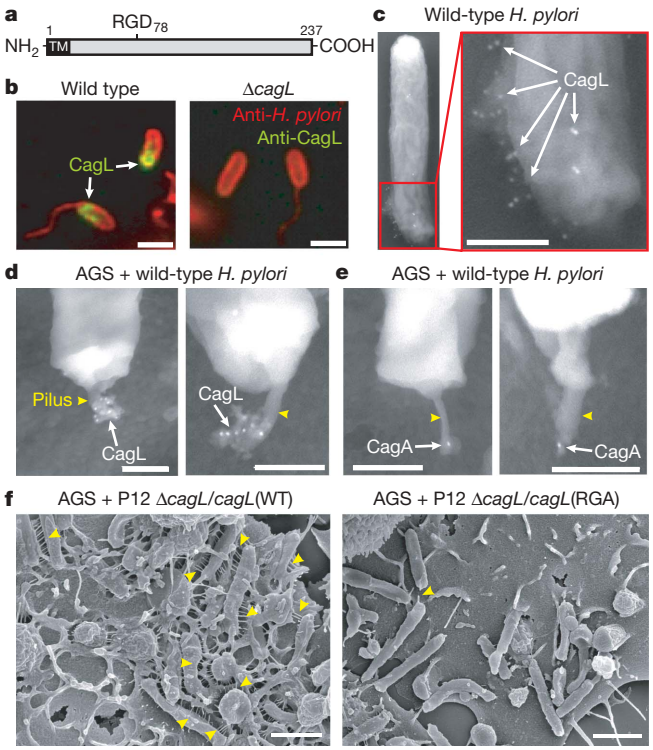


Figure 2 | The surface protein CagL associates with the type-IV secretion pilus surface during infection. **a**, Schematic of CagL showing the predicted leader peptide (TM) and RGD motif (residues 76–78). **b**, Wild-type (WT) *H. pylori* and the Δ*cagL* deletion mutant were immunostained without membrane permeabilization. Scale bars, 2 μm. **c**, Wild-type *H. pylori* labelled with anti-CagL coupled with gold particles (arrows). Scale bar, 250 nm. **d**, **e**, Wild-type *H. pylori* infecting AGS cells incubated with anti-CagL (**d**) or anti-CagA (**e**) coupled with gold particles (arrows) show distinct labelling of type-IV-secretion-dependent pili (arrowheads) in two representative examples. Scale bars, 250 nm. **f**, FESEM of wild-type *H. pylori* and the *cagL*(RGA) mutant at 2 h after infection of AGS cells. Scale bars, 2 μm.

type-IV secretion and phosphorylation of CagA. They rule out the scenario where CagA is translocated by an RGD-motif-independent mechanism and the RGD peptide merely induces phosphorylation of the translocated CagA.

CagL binds integrin $\alpha_5\beta_1$ in an RGD-dependent manner

Recombinant wild-type CagL bound to purified integrin $\alpha_5\beta_1$ with a higher affinity (dissociation constant (K_d) of 0.09 ± 0.02 μM) and higher association and dissociation rates as compared with mutant CagL(RGA) (K_d of 0.36 ± 0.11 μM) (Fig. 4a and Supplementary Figs 6 and 7). These observations suggest that CagL is able to interact directly with purified integrin $\alpha_5\beta_1$ *in vitro* in an RGD-motif-specific manner. The affinities and specificities of integrin–ligand interactions *in vivo*, in contrast to those *in vitro*, are influenced by factors such as inside-out signalling, integrin clustering and receptor crosstalk[1–4]. Indeed, CagL bound to cell-surface-expressed integrin $\alpha_5\beta_1$ with even greater affinity and specificity as compared with the *in vitro* interaction, as indicated by the following observations. First, latex beads coated with CagL bound abundantly to AGS cells but only poorly to integrin-β_1 knockdown cells (Fig. 4b and Supplementary Fig. 8a–c). Second, the binding of CagL(RGA)-coated latex beads to AGS cells was significantly reduced as compared to wild-type CagL (Supplementary Fig. 8a–c). Third, CagL-coated beads co-localized with integrin β_1 on binding to AGS cells (Supplementary Fig. 9a), whereas the integrin β_1 signal was often weak or undetectable with CagL(RGA)-coated beads (Supplementary Fig. 9a, right panel). Fourth, the function-blocking antibodies BIIG2 and AIIB2 against $\alpha_5\beta_1$ integrins as well as the CagL-derived RGD peptide, but not the control RAD peptide, inhibited binding of CagL to AGS cells (Supplementary Fig. 9b, c). Fifth, invasion of *E. coli* HB101 *invA*$^+$ into AGS cells, which requires integrin β_1 (ref. 6), was blocked by pre-infection of AGS cells with wild-type *H. pylori*, but not with the *cagL* deletion mutant (Δ*cagL*) or the *cagL*(RGA) mutant (Supplementary Fig. 10). These various findings show conclusively that CagL uses its RGD motif to interact with the surface receptor integrin $\alpha_5\beta_1$ on eukaryotic host cells, and support the hypothesis that CagL triggers type-IV secretion and phosphorylation of CagA through direct interaction with integrin $\alpha_5\beta_1$.

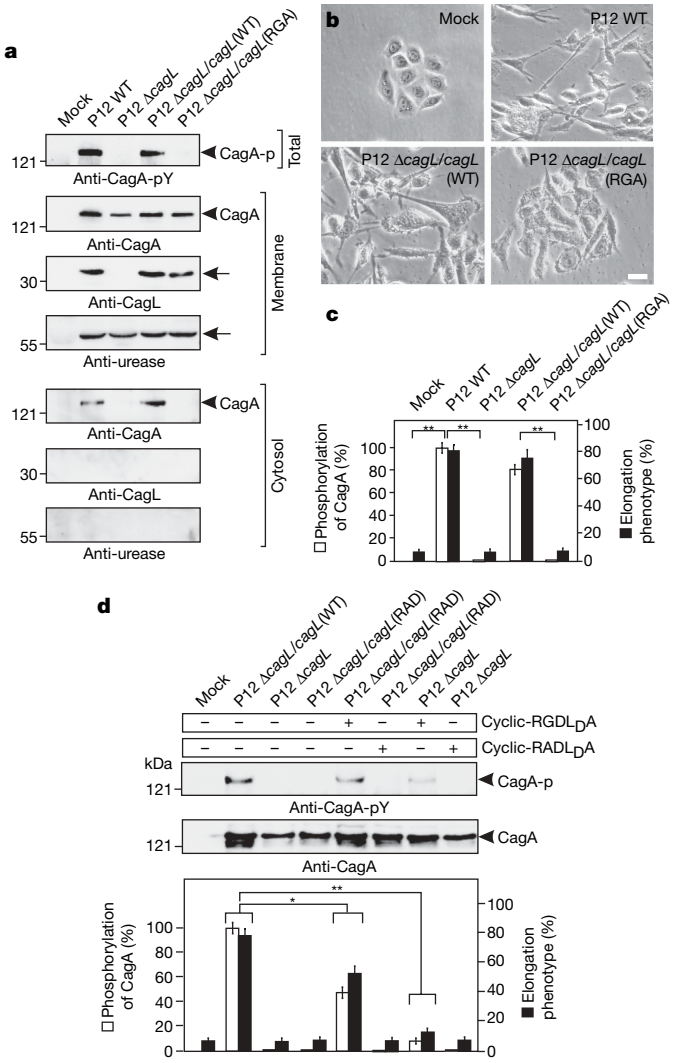


Figure 3 | *H. pylori* CagL and its RGD motif are essential for type-IV secretion of CagA. **a**, AGS cells were infected with *H. pylori* P12 wild type and various *cagL* mutant strains. The respective membrane and cytosol fractions were immunoblotted with the antibodies indicated. Most CagA and urease in the membrane fraction originated from attached bacteria that could not be completely removed during fractionation[13]. The lower anti-urease and anti-CagL blots show that the cytosolic fractions are not contaminated with *H. pylori* proteins[14]. **b**, Morphologies of AGS cells infected with various *H. pylori* strains as indicated. Scale bar, 10 μm. **c**, The corresponding percentage of elongated cells and relative amounts of phosphorylated CagA are shown. **d**, AGS cells were infected with various *H. pylori* strains as indicated. The RGD peptide (see Methods) partially complemented the function of CagL(RAD) but only weakly the function of the Δ*cagL* mutant during infection. Tyrosine phosphorylation of CagA in the total cell lysate was analysed by immunoblotting. Percentages of elongated cells and relative amounts of phosphorylated CagA are shown (graph). Error bars in **c** and **d** indicate ±s.d. ($n = 3$); asterisk, $P \leq 0.01$; double asterisk, $P \leq 0.001$.

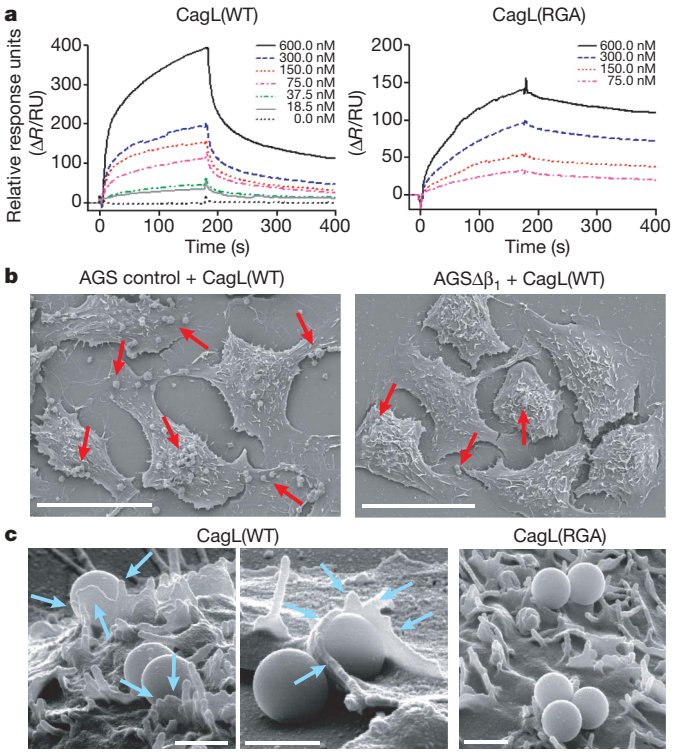


Figure 4 | RGD-dependent interaction of CagL with integrin $\alpha_5\beta_1$. **a**, Surface plasmon resonance sensograms of the interaction of CagL(WT) (left) and CagL(RGA) (right) (0–600 nM) with immobilized integrin $\alpha_5\beta_1$. ΔR quantifies the response between the measurement and reference flow cells given in resonance units (RU). **b**, FESEM of CagL(WT)-coated beads binding to AGS control and AGSΔβ_1 cells. Arrows indicate examples of bound beads. Scale bars, 20 μm. **c**, FESEM of CagL(WT)-coated or CagL(RGA)-coated beads binding to AGS cells at high magnification. Arrows indicate examples of membrane ruffling and engulfment induced by CagL(WT)-coated beads. Scale bars, 1 μm.

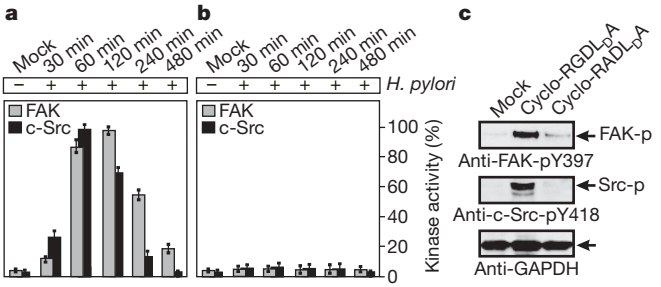


Figure 5 | The RGD motif of CagL is crucial for activation of FAK and Src tyrosine kinases. **a**, **b**, Kinase activities of FAK and Src during infection of AGS cells with *H. pylori* P12 expressing either wild-type CagL (**a**) or CagL(RAD) mutant (**b**) were examined at the time points indicated. Kinase activity was determined by immunoblotting with anti-FAK-pY397 and anti-Src-pY418 antibodies, respectively (the corresponding blots are shown in Supplementary Fig. 11). **c**, Immunoblots of AGS cells treated for 12 h with the indicated CagL-derived RGD or RAD peptides, respectively (see Methods). The results show that the RGD motif of CagL is sufficient to induce FAK and Src activities.

CagL activates FAK and Src tyrosine kinases

Analyses of bound CagL-coated latex beads on integrin-β_1-positive cells at higher resolution showed that CagL beads triggered membrane ruffles on the host cell surface as well as intimate contact with the host cell membrane, whereas beads coated with the CagL(RGA) mutant exhibited only limited contact with the host surface (Fig. 4c, arrows). These observations support the view that CagL functions as a specialized adhesin that not only anchors the type-IV secretion apparatus to the host surface through binding to integrin, but also promotes signal transduction. These features of CagL complement the fundamental bacterial adherence functions of the two canonical *H. pylori* adhesins BabA and SabA[17,18] during *H. pylori* infections. To understand further the downstream signalling of the CagL–integrin interaction, we examined the ability of CagL to trigger activation of FAK, as the latter is a key sensor of integrin engagement. On activation, FAK is autophosphorylated at tyrosine residue Y397, which then serves as a high-affinity binding site for the SH2 domain of Src, resulting in an active FAK–Src signalling complex[2]. Our results show that both the infection of AGS cells with *H. pylori* expressing wild-type CagL and incubation of AGS cells with the RGD peptide led to transient activation of FAK and Src, as indicated by the accumulation of phosphorylated FAK-pY397 and Src-pY418, respectively (Fig. 5a, c and Supplementary Fig. 11a). The activation of FAK and

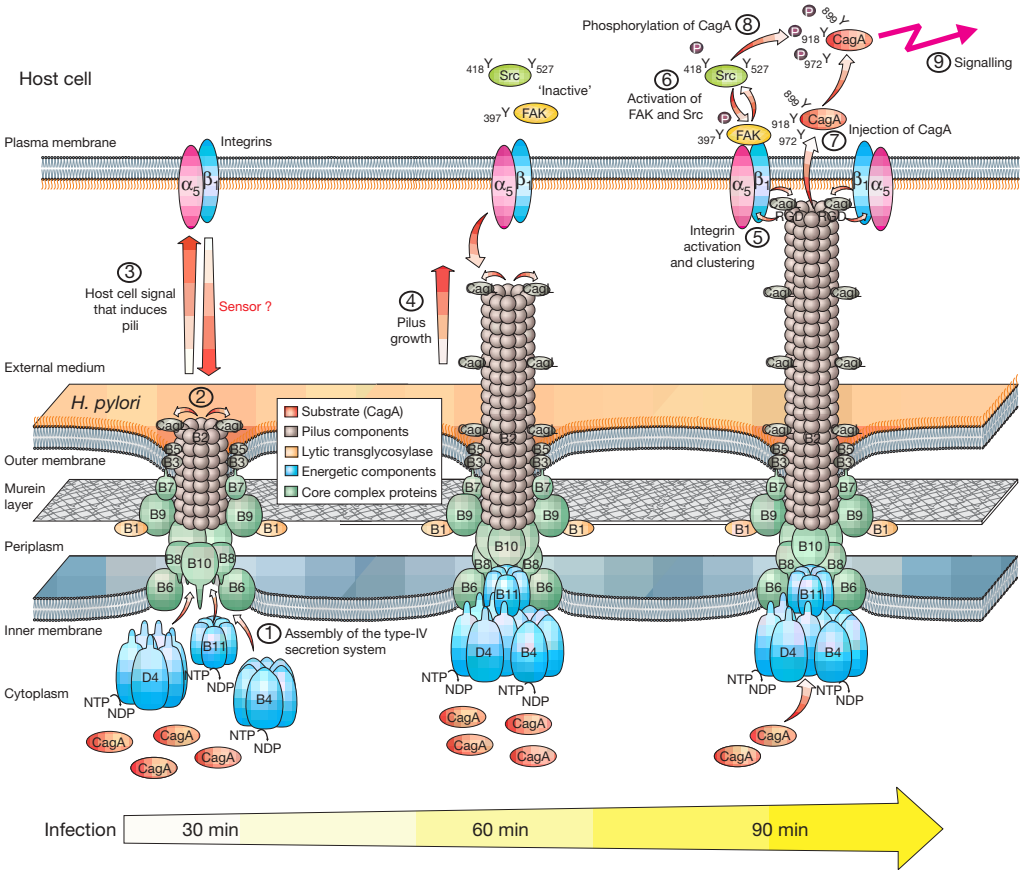


Figure 6 | Three-step model for the activation of the *H. pylori* type-IV secretion system and phosphorylation of CagA through interaction of CagL with integrin. The illustrated type-IV secretion assembly mechanism is based on homology to the prototypical agrobacterial system, which is composed of 11 VirB proteins (B1–11) and the so-called coupling protein VirD4 (D4)[8,9,12], as indicated. The *H. pylori* type-IV secretion apparatus consists of up to 32 proteins including VirB/D homologues and proteins of as-yet-unknown functions such as CagL[12]. In the absence of host contact, CagL is delivered to the bacterial surface or even specifically to the surface of the type-IV secretion pilus (steps 1–2). During infection, *H. pylori* establishes contact with the host cell. Surface-located CagL then binds to the transmembrane receptor integrin $\alpha_5\beta_1$ (step 3). This interaction stimulates pili elongation by an unknown mechanism (step 4). Meanwhile, binding of pilus-located CagL through its RGD motif to host integrin $\alpha_5\beta_1$ activates this receptor (step 5). Activated integrin $\alpha_5\beta_1$ triggers downstream signalling, leading to activation/phosphorylation of both FAK and Src kinases (step 6), which subsequently induce actin-cytoskeletal rearrangements, plasma-membrane dynamics and possibly clustering of integrins and activation of other receptors. These changes are important with regard to injecting CagA from the pilus into the host cell cytoplasm (step 7). Finally, activated Src mediates tyrosine phosphorylation of translocated CagA (step 8), which then stimulates downstream signalling, leading to cell elongation or induction of IL-8 (step 9).

Src coincided with the occurrence of CagA-pY in infected cells (Supplementary Fig. 11a), indicating that FAK activation and subsequently the activation of Src rapidly results in tyrosine phosphorylation of CagA. In contrast, activation of FAK and Src was seen neither with the *H. pylori* CagL(RGA) mutant nor the RAD peptide control (Fig. 5b, c and Supplementary Fig. 11b), indicating that the RGD motif of CagL has a crucial role in the activation of FAK and Src.

Discussion

Taken together, we have shown that CagL is able to act bifunctionally to hijack host integrin receptors for injection of virulence factors into eukaryotic cells as well as to activate host tyrosine kinases. We propose that this elegant mechanism allows CagL to not only promote receptor-dependent translocation of CagA but also activate concomitantly the tyrosine phosphorylation of CagA at the site of injection (Fig. 6). Whether the CagL-mediated signalling might facilitate type-IV secretion by altering membrane dynamics is an intriguing proposition to be investigated further. Notably, the observation that both integrin β_1 function-blocking antibody and knockdown of integrin β_1 markedly reduced the number of elongated *H. pylori* pili (Supplementary Fig. 4e, f) suggests that interaction of integrin β_1 with CagL is important for the formation of visible type-IV secretion pili (Fig. 6). Furthermore, the CagL-derived RGD peptide only partially complemented the *H. pylori cagL*(RAD) mutant in the induction of type-IV secretion and phosphorylation of CagA, and failed to complement the *ΔcagL* deletion mutant in these functions (Fig. 3d), indicating that a direct contact between integrin $\alpha_5\beta_1$ and CagL on *H. pylori* is likely to be essential for full type-IV secretion function, and that other amino acids apart from the RGD motif in CagL also have a function in activating type-IV secretion and phosphorylation of CagA.

We have demonstrated a previously unidentified role of integrin $\alpha_5\beta_1$ in microbial pathogenesis, making integrin $\alpha_5\beta_1$ the first human receptor shown to be exploited by a human bacterial pathogen for type-IV secretion of virulence factors into mammalian cells. Our findings support an intriguing mechanism by which the *H. pylori* type-IV secretion system, through activation of integrin $\alpha_5\beta_1$ by the pilus-located adhesin and sensor molecule CagL, ensures phosphorylation of translocated CagA directly at the site of injection at focal adhesions and subsequently promotes downstream signalling in the host cell. Given that integrin $\alpha_5\beta_1$ preferentially localizes to the basolateral surface of polarized epithelial cells, we propose that *H. pylori* disrupts cell–cell junctions and invades between cells to come into contact with integrins[10,19]. Although this study provides clear evidence that $\alpha_5\beta_1$ integrin is a principal molecule in the interaction with CagL and type-IV secretion of CagA, the data presented do not preclude the possibility that other integrin receptors expressed on AGS cells might potentially also be involved in the binding of CagL to AGS cells. However, the notion that the type-IV secretion system functions through specific recognition of a host cell receptor adds a new dimension to our understanding of *H. pylori* pathogenesis, and opens up new opportunities for investigating type-IV-secretion-mediated protein translocation in general. The fact that CagL activates host tyrosine kinases and engages integrin $\alpha_5\beta_1$ for triggering type-IV secretion provides new insights into the molecular basis of *H.-pylori*-induced malignancy. CagL might be a novel drug target for combating the severe gastric diseases caused by *cag*PAI-positive *H. pylori* strains.

METHODS SUMMARY

The Methods and Supplementary Information provide detailed information regarding all experimental procedures: a list of host cells, bacteria, infection and phosphorylation assays; synthetic peptides, antibodies and western blotting; overexpression and purification of CagL proteins; cell attachment assays and binding of ligand-coated latex beads; binding of CagL to integrin $\alpha_5\beta_1$ determined by surface plasmon resonance; immunofluorescence staining and confocal laser scanning microscopy; field and immunofield emission scanning electron microscopy; and statistical analysis.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 May; accepted 21 August 2007.

1. Hynes, R. O. Integrins: bidirectional, allosteric signaling machines. *Cell* **110,** 673–687 (2002).
2. Mitra, S. K. & Schlaepfer, D. D. Integrin-regulated FAK-Src signaling in normal and cancer cells. *Curr. Opin. Cell Biol.* **18,** 516–523 (2006).
3. Ruoslahti, E. RGD and other recognition sequences for integrins. *Annu. Rev. Cell Dev. Biol.* **12,** 697–715 (1996).
4. Eble, J. A. & Kühn, K. (eds) *Integrin-Ligand Interaction* (Springer, Heidelberg, 1997).
5. Isberg, R. R. & Tran Van Nhieu, G. Binding and internalization of microorganisms by integrin-receptors. *Trends Microbiol.* **2,** 10–14 (1994).
6. Isberg, R. R. & Leong, J. M. Multiple β1 chain integrins are receptors for invasin, a protein that promotes bacterial penetration into mammalian cells. *Cell* **60,** 861–871 (1990).
7. Boyle, E. C. & Finlay, B. B. Bacterial pathogenesis: exploiting cellular adherence. *Curr. Opin. Cell Biol.* **15,** 633–639 (2003).
8. Cascales, E. & Christie, P. J. The versatile bacterial type IV-secretion systems. *Nature Rev. Microbiol.* **1,** 137–149 (2003).
9. Backert, S. & Meyer, T. F. Type IV secretion systems and their effectors in bacterial pathogenesis. *Curr. Opin. Microbiol.* **9,** 207–217 (2006).
10. Amieva, M. R. *et al.* Disruption of the epithelial apical-junctional complex by *Helicobacter pylori* CagA. *Science* **300,** 1430–1434 (2004).
11. Peek, R. M. Jr & Blaser, M. J. *Helicobacter pylori* and gastrointestinal tract adenocarcinomas. *Nature Rev. Cancer* **2,** 28–36 (2002).
12. Covacci, A. & Rappuoli, R. Tyrosine-phosphorylated bacterial proteins: Trojan horses for the host cell. *J. Exp. Med.* **191,** 587–592 (2000).
13. Hatakeyama, M. The role of *Helicobacter pylori* CagA in gastric carcinogenesis. *Int. J. Hematol.* **84,** 301–308 (2006).
14. Backert, S. *et al.* Phosphorylation of tyrosine 972 of the *Helicobacter pylori* CagA protein is essential for induction of a scattering phenotype in gastric epithelial cells. *Mol. Microbiol.* **42,** 631–644 (2001).
15. Wennerberg, K. *et al.* β1 integrin-dependent and-independent polymerization of fibronectin. *J. Cell Biol.* **132,** 227–238 (1996).
16. Rohde, M. *et al.* A novel sheathed surface organelle of the *H .pylori cag*-type IV-secretion system. *Mol. Microbiol.* **49,** 219–234 (2003).
17. Mahdavi, J. *et al. Helicobacter pylori* SabA-adhesin in persistent infection and chronic inflammation. *Science* **297,** 573–578 (2002).
18. Aspholm-Hurtig, M. *et al.* Functional adaptation of BabA, the *H. pylori* ABO-blood group antigen binding adhesin. *Science* **305,** 519–522 (2004).
19. Papini, E. *et al.* Selective increase of the permeability of polarized epithelial cell monolayers by *Helicobacter pylori* vacuolating toxin. *J. Clin. Invest.* **102,** 813–820 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank I. Schleicher, D. Schmidt and N. Tegtmeyer for technical assistance, and S. Johansson and R. Faessler for providing GD25 cells. U. Strijowski synthesized the peptides. This work was supported through Magdeburger-Forschungsverbund and a DFG priority program grant.

Author Contributions T.K. purified CagL and performed cell binding studies, flow cytometry and integrin expression assays; S.B. designed the concept, performed the cloning work, constructed *H. pylori* mutants and supervised the project; D.Z. did the infections experiments, cell fractionation and western blotting; S.U. and N.S. performed the Biacore binding studies and designed the synthetic CagL peptides; M.R. and J.B. performed the FESEM studies; R.H. and S.W. conducted the immunofluorescence experiments; R.M. did the CD analyses and W.K. provided materials; T.K. and S.B. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.B. (Steffen.Backert@med.ovgu.de).

METHODS

Host cells, bacteria, infection and phosphorylation assays. The human gastric adenocarcinoma cell line AGS[14,20], the integrin β_1-deficient GD25 cells and GD25 cells re-expressing integrin β_1A (GD25β1A) have been described[15]. AGS cells stably expressing short hairpin RNA (shRNA) targeting integrin β_1 RNA (AGSΔβ_1 knockdown cells) or control cells expressing shRNA specific for firefly luciferase-GL2 RNA were generated by retroviral transduction. Knockdown of integrin β_1 was verified using standard flow cytometry (FACS-Calibur, BD Biosciences) and immunoblotting. *H. pylori* strains P1, P12 and their isogenic mutants P1 Δ*cagL*, P1 Δ*flaA*, P1 Δ*virB7*, P1 Δ*virB10*, P1 Δ*virB11*, P1 Δ*cagA*/*cagA*(WT), P1 Δ*cagA*/*cagA*(Y972F) or P12 Δ*cag*PAI have been reported previously[14,20]. Cloning, mutagenesis and genetic complementation of *cagL* was done using standard procedures[14]. Cell monolayers were infected with *H. pylori* as described[14,20]. The number of infected AGS cells exhibiting elongation phenotype was quantified in ten different 0.25-mm^2 fields. To determine the number of intracellular bacteria, gentamicin protection assays were performed[20]. The number of intracellular bacteria (colony-forming units, c.f.u.) was obtained in three independent experiments. The preparation of membrane/nuclear and cytoplasmic fractions of infected AGS cells has been described previously[14]. *In vitro* phosphorylation assays with recombinant Src (Upstate) was performed as described previously[21].

Synthetic peptides, antibodies and western blotting. The cyclic peptides were synthesized as previously described[22]. The following antibodies were used: rabbit polyclonal anti-*H. pylori* antibody (Biomeda); AIIB2 and BIIG2 specific for β_1 and α_5 integrin, respectively (Developmental Studies Hybridoma Bank); anti-integrin $\alpha_v\beta_3$ (clone LM609, Chemicon); monoclonal anti-CagA antibody (AustralBiologicals); FITC-conjugated mouse anti-human integrin β_1, rabbit anti-human integrin β_1 antibody and monoclonal anti-vinculin (all from Santa Cruz). The mouse anti-CagL, rabbit anti-CagA and anti-urease antisera against the purified proteins were prepared by Biogenes. The rabbit anti-CagA-pY972 antibody was produced using a peptide-based antigen (H-VGLSASPEPIpYAT, Jerini). Western blotting was performed as described[14,20]. To examine the type-IV secretion and phosphorylation of CagA during infections, total lysate of infected cells or the corresponding membrane and cytosol fractions were analysed by immunoblotting with anti-pY99 (Santa Cruz). Blots were probed sequentially with anti-CagA and anti-CagL antibodies to ensure equal amounts of bacteria in each lane. To rule out the possibility that the cytosolic fractions were contaminated with *H. pylori* proteins, all blots were probed with anti-urease[14]. The presence of urease in the membrane fractions of infected AGS cells is due to attached *H. pylori*. Band intensities were quantified with the Lumi-Imager F1 (Roche Diagnostics). Images were processed for contrast and brightness using Adobe Photoshop (version 6.0).

Overexpression and purification of CagL proteins. To construct the vectors for overexpression of wild-type CagL or CagL(RGA) in *E. coli*, a DNA fragment corresponding to amino acid residues 21–237 of the protein (minus the predicted signal peptide) was amplified by PCR and ligated into pET-28a vector. CagL proteins were purified as follows: *E. coli* BL21(DE3) transformed with the corresponding overexpression plasmids were grown in LB medium at 37 °C for 2 h with shaking. Bacterial pellets were re-suspended in ice-cold buffer CW (50 mM KH_2PO_4-K_2HPO_4, pH 7.5, 200 mM NaCl) supplemented with protease inhibitor cocktail (Roche). After sonication, the overexpressed CagL present in the inclusion bodies was solubilized in buffer LW (50 mM KH_2PO_4-K_2HPO_4, pH 7.5, 200 mM NaCl, 6 M guanidine hydrochloride) and refolded in ice-cold refolding buffer (52 mM Tris-HCl, pH 8.2, 20 mM NaCl, 834 μM KCl, 1.1 mM EDTA, 2.1 mM reduced glutathione, 210 μM oxidized glutathione). After refolding, CagL was further purified by metal-chelate affinity chromatography through Talon resin (BD Biosciences) and gel filtration in buffer CW through Sephacryl S-200 (16/60, Amersham Biosciences) to >95% homogeneity. The folded conformations of the purified CagL and CagL(RGA) were confirmed by circular dichroism[23] using a J-720 spectropolarimeter (Jasco Instruments).

Cell attachment assays and binding of ligand-coated latex beads. Cell attachment assays were performed according to procedures described previously[24]. Latex beads (1.1 μm in diameter, Sigma) were coated with CagL or CagL(RGA) mutant protein as follows: the beads were incubated with protein ligand (1 mg ml^{-1}) in coupling buffer (50 mM MES pH 6.1, 200 mM NaCl) at 4 °C overnight. Beads were then washed with coupling buffer and nonspecific sites were blocked with 2% BSA for 1 h. AGS cell monolayer was co-incubated with protein-coated latex beads at an approximate cell:bead ratio of 1:100 for 3 h. After washing with RPMI, cells were fixed with 3.8% paraformaldehyde.

Binding of CagL to integrin $\alpha_5\beta_1$ determined by surface plasmon resonance. Purified integrin $\alpha_5\beta_1$ (Chemicon) was immobilized by amino coupling to the surface of a CM5 carboxymethyl dextran sensorchip. The analysis was performed as described previously[22]. Data were processed using Biaevaluation software (version 4.1).

Immunofluorescence staining and confocal laser scanning microscopy. Immunofluorescence staining with different antibodies as shown in each experiment was performed as described[14,20]. Samples were analysed using a Leica TCS SP2 confocal laser scanning microscope system equipped with a DM-IRE2 microscope and different lasers (Leica Microsystems). To avoid spectral overlap and channel crosstalk, FITC, TRITC, CY5 and Alexa-350 fluorophores were excited sequentially with argon-laser (488 nm), green helium (543 nm), red helium (633 nm) and ultraviolet laser (364 nm). DAPI (4,6-diamidino-2-phenylindole) staining was used to visualize DNA.

Field and immunofield emission scanning electron microscopy (FESEM). Procedures for the FESEM of *H. pylori* and *H.-pylori*-infected AGS cells were carried out as described previously[16]. For immuno-FESEM of CagL, *H. pylori* samples were incubated with rabbit anti-CagL IgG antibodies (100 μg IgG protein per ml) followed by incubation with 15-nm protein-A-coated gold particles as described[16]. *H.-pylori*-infected AGS samples on cover slips were incubated with rabbit anti-CagL IgG antibodies (100 μg IgG protein per ml) for 1 h at 30 °C, followed by washing in PBS and incubation with 15-nm protein-A-coated gold particles, or samples were incubated with mouse anti-CagA antibodies followed by 15-nm anti-mouse IgG-coated gold particles. All samples were coated with a thin carbon film.

Statistical analysis. All data were evaluated using the Student's *t*-test with SigmaStat statistical software (version 2.0). Statistical significance was defined by $P \leq 0.01$ (asterisk) and $P \leq 0.001$ (double asterisk).

20. Kwok, T. *et al.* Specific entry of *Helicobacter pylori* into cultured gastric epithelial cells *via* a zipper-like mechanism. *Infect. Immun.* **70**, 2108–2120 (2002).
21. Selbach, M. *et al.* Src is the kinase of the *Helicobacter pylori* CagA protein *in vitro* and *in vivo*. *J. Biol. Chem.* **277**, 6775–6778 (2002).
22. Zimmermann, D. *et al.* Integrin $\alpha_5\beta_1$ antagonists: Biological evaluation using cell adhesion assay and surface plasmon resonance. *ChemBioChem* **6**, 272–276 (2005).
23. Johnson, W. C. Jr. Protein secondary structure and circular dichroism: a practical guide. *Proteins* **7**, 205–214 (1990).
24. Steffensen, B. *et al.* The involvement of the fibronectin type II-like modules of human gelatinase A in cell surface localization and activation. *J. Biol. Chem.* **273**, 20622–20628 (1998).

ARTICLES

Protein-based peptide-bond formation by aminoacyl-tRNA protein transferase

Kazunori Watanabe[1]*, Yukimatsu Toh[1]*, Kyoko Suto[1], Yoshihiro Shimizu[2], Natsuhisa Oka[2], Takeshi Wada[2] & Kozo Tomita[1]

Eubacterial leucyl/phenylalanyl-tRNA protein transferase (LF-transferase) catalyses peptide-bond formation by using Leu-tRNA^{Leu} (or Phe-tRNA^{Phe}) and an amino-terminal Arg (or Lys) of a protein, as donor and acceptor substrates, respectively. However, the catalytic mechanism of peptide-bond formation by LF-transferase remained obscure. Here we determine the structures of complexes of LF-transferase and phenylalanyl adenosine, with and without a short peptide bearing an N-terminal Arg. Combining the two separate structures into one structure as well as mutation studies reveal the mechanism for peptide-bond formation by LF-transferase. The electron relay from Asp 186 to Gln 188 helps Gln 188 to attract a proton from the α-amino group of the N-terminal Arg of the acceptor peptide. This generates the attacking nucleophile for the carbonyl carbon of the aminoacyl bond of the aminoacyl-tRNA, thus facilitating peptide-bond formation. The protein-based mechanism for peptide-bond formation by LF-transferase is similar to the reverse reaction of the acylation step observed in the peptide hydrolysis reaction by serine proteases.

Regulated degradation of intracellular proteins is essential for the control of post-translational gene expression in all organisms. The N-end rule pathway, one of the primary proteolytic pathways, controls the half-lives of proteins by destroying them according to their N-terminal amino acid residue[1–3], and is involved in many cellular functions, such as chromosomal segregation fidelity, apoptosis regulation and nitric oxide detection[4–6]. In eukaryotes, proteins with N-terminal primary destabilizing residues, such as basic (Arg, Lys or His) or bulky and hydrophobic (Phe, Leu, Trp, Tyr or Ile) amino acid residues, are subjected to poly-ubiquitylation by multiple E3 ligases, which is followed by degradation mediated by the 26S proteasome complexes[7,8]. In eubacteria, proteins with primary destabilizing residues, such as bulky and hydrophobic amino acid residues (Leu, Phe, Trp or Tyr), are recognized by the adaptor protein ClpS, and thereafter are degraded by the proteasome-like protease ClpAP[9,10]. In the degradation of proteins that is governed by the N-end rule, an aminoacyl-transfer(t)RNA protein transferase is involved in the conjugation of a primary destabilizing amino acid to the N-terminal residue of proteins, using cognate aminoacyl-tRNAs as amino acid donor substrates[11]. The aminoacyl-tRNA protein transferase is a member of the FemABX family[12–14]. The FemABX family contains a GCN5-related *N*-acetyltransferase fold and catalyses the same chemical reaction[15], in which the amino acid is transferred from an aminoacyl-tRNA to an amino group of a protein, and thus facilitates peptide-bond formation in a ribosome-independent manner. Although crystallographic analyses of FemABX family proteins have been reported[13,14], the catalytic mechanism of peptide-bond formation by this family of proteins remains obscure. Eubacterial leucyl/phenylalanyl-tRNA protein transferase (LF-transferase) catalyses peptide-bond formation by using Leu-tRNA^{Leu} (or Phe-tRNA^{Phe}) and a protein bearing an N-terminal Arg (or Lys) as donor and acceptor substrates, respectively[16–18] (Fig. 1a). Here we analysed the crystal structures of the *Escherichia coli* LF-transferase complex with phenyalanyl adenosine (rA-Phe), with or without a short peptide bearing an N-terminal Arg residue. In the presence of both the donor and acceptor substrates, the peptide formation proceeded within the crystals, and the product peptide bearing Phe at the N terminus was retained on the LF-transferase. Combining the two separate structures, one with rA-Phe and one with the product peptide, into one structure has revealed the mechanism underlying the substrate specificity and the catalytic reaction of peptide-bond formation by LF-transferase (Supplementary Fig. 1).

Recognition of the donor substrate

Recent crystallographic structural analyses of *E. coli* LF-transferase and its complex with puromycin revealed that the *p*-methoxybenzyl group of puromycin is accommodated in a highly hydrophobic pocket, with a shape and size suitable for hydrophobic amino acid residues that lack a branched β-carbon, such as leucine and phenylalanine[19]. However, the mechanisms for the recognition of the acceptor protein bearing N-terminal Arg or Lys, and for the catalytic reaction of peptide-bond formation by LF-transferase have remained unsolved. An attempt to capture the ternary reaction complex of LF-transferase, puromycin and a peptide bearing N-terminal Arg was unsuccessful[19]; this might have been due to the different chemical structure of puromycin from that of the 3′-end of an aminoacyl-tRNA.

Phenylalanyl-tRNA synthetase is known to aminoacylate the 2′-hydroxyl group of the 3′-end A76 of tRNA^{Phe} (ref. 20). However, an *in vitro* analysis using phenylalanylated tRNA^{Phe} with 3′-deoxyl adenosine at position 76 showed that LF-transferase cannot transfer phenylalanine to the N-terminal Arg of α-casein efficiently (Supplementary Fig. 2), suggesting that LF-transferase uses transacylated 3′-phenylalanylated tRNA^{Phe} as a substrate. Therefore, the complex structure of LF-transferase with rA-Phe, in which phenylalanine is linked to the 3′-hydroxyl of adenosine through an ester bond, was analysed (Supplementary Table 1).

[1]Institute of Biological Resources and Functions, National Institute of Advanced Industrial Sciences and Technology, 1-1-1, Higashi, Tsukuba-shi, Ibaraki 305-8566, Japan. [2]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8562, Japan.
*These authors contributed equally to this work.

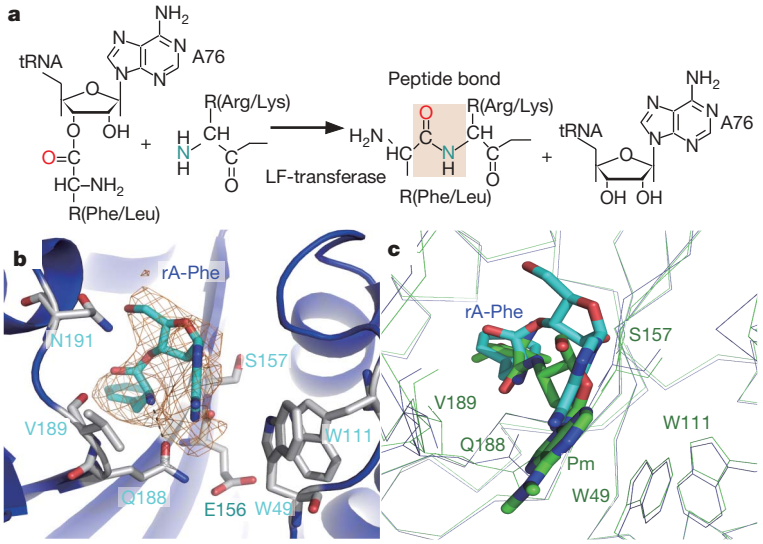


Figure 1 | Recognition of the aminoacyl-tRNA analogue, phenylalanyl adenosine. **a**, Peptide-bond formation catalysed by LF-transferase. **b**, Structure of the LF-transferase complex, with rA-Phe shown as a cyan-coloured stick model. σ_a-weighted simulated annealing F_O–F_C omit electron density map of rA-Phe contoured at 3.0 σ, coloured orange. **c**, Overlay of two complex structures of LF-transferase, with rA-Phe and with puromycin (Pm). rA-Phe and puromycin are shown as stick models and are coloured cyan and green, respectively. The root mean squared (r.m.s.) deviation of the ribose group is 3.8 Å. R, side-chain residue.

In the complex structure of LF-transferase and rA-Phe, rA-Phe is located at almost the same position in LF-transferase as puromycin (Fig. 1b and c). The electron density of the adenosine of rA-Phe is not completely occupied in the final $2F_O$–F_C map, as compared with that of the benzyl group of rA-Phe, and the B-factor of adenosine is higher than that of the benzyl group, indicating that the adenosine of rA-Phe is relatively mobile (Supplementary Fig. 3). This might partly reflect the fact that rA-Phe is a mixture of isomers of 2′- and 3′-aminoacylated adenosine (ratio is 1:1), generated by spontaneous intramolecular transacylation, and is sensitive to spontaneous hydrolysis under the crystallization conditions. (Supplementary Fig. 3).

In the complex structure with rA-Phe, the benzyl group of the analogue is accommodated within the C-shaped hydrophobic pocket, as is also observed in the complex with puromycin[19]. The ribose of rA-Phe in the complex structures adopts a C_3' endo conformation, as the conventional ribose of RNA does, and the 2′-hydroxyl group of the ribose forms a hydrogen bond with the Oγ atom of Ser 157. The α-amino group of rA-Phe forms a hydrogen bond with the Oε atom of Gln 188, and the Nε atom of Gln 188 forms a hydrogen bond with the Oε atom of Glu 156, creating an intermolecular hydrogen network. The adenine base stacks with Trp 49 and interacts with Val 189 weakly, and no specific hydrogen bonds are observed. The conformation of rA-Phe bound to LF-transferase differs from that of puromycin (Fig. 1c). Although the benzyl group of rA-Phe and the *p*-methoxybenzyl group of puromycin are accommodated within the same hydrophobic pocket, owing to the different chemical structures of these analogues, the ribose groups of these analogues do not superpose completely.

Recognition of the acceptor protein

A short peptide bearing an N-terminal Arg (α-casein fragment; RYLGYL) can be a good acceptor substrate for both phenylalanyl and leucyl transfer from the cognate aminoacyl-tRNA by LF-transferase[19]. The crystals of LF-transferase were soaked in a solution containing both rA-Phe and the α-casein fragment, and the complex structure was analysed. The final $2F_O$–F_C map clearly showed the electron density corresponding to the N-terminal part of the α-casein fragment (amino acid residues 1–4; RYLG), whereas that of the carboxy-terminal two amino acids (amino acid residues 5 and 6; YL) was not observed (Fig. 2a). An additional electron density continuous with the N-terminal Arg of the α-casein fragment was observed in the C-shaped hydrophobic pocket. An *in vitro* analysis revealed that the Phe of rA-Phe could be transferred to the α-casein fragment (Supplementary Fig. 4), indicating that rA-Phe can be a minimal amino acid donor substrate of LF-transferase. Therefore, the additional electron density observed was assigned as that of Phe, which was transferred to the N-terminal Arg of α-casein fragment from rA-Phe in the crystals. The electron density of adenosine—the by-product of the reaction—is not visible in the structure. This suggested that the structure represents the reaction state when peptide-bond formation is complete, with the product peptide ready for release from the enzyme. Consistent with this, when the crystals were soaked in a solution containing both the peptide bearing an N-terminal Phe (FRYLG) and adenosine, only the electron density of the peptide was observed. The conformations of the product peptide and the peptide bearing the N-terminal Phe in both LF-transferase structures superposed well (Supplementary Fig. 5).

The body of the product peptide (FRYLG), except for the N-terminal Phe and Arg, is recognized in a sequence-independent manner by LF-transferase (Fig. 2a). The main-chain amide of Arg and that of Tyr of the peptide electrostatically interact with the

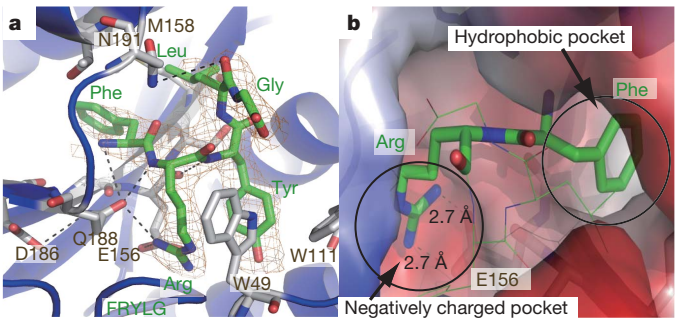


Figure 2 | Recognition of the product-peptide bearing an N-terminal Phe. **a**, Structure of the LF-transferase complex with a product peptide. The peptide FRYLG is shown as a green-coloured stick model. Final $2F_O$–F_C electron density map of the product peptide, contoured at 1.0 σ, coloured orange. The amino acid residues of the product peptide are depicted by the three letter code, coloured green. **b**, Electrostatic potential of LF-transferase. Blue and red represent positively and negatively charged areas, respectively. The Phe and Arg residues of the product peptide are shown in stick models, and the others are shown in green-coloured line models.

main-chain carbonyl of Glu 156. The carbonyl of Leu in the peptide moderately interacts with the Nδ atom of Asn 191. The benzyl group of Phe, transferred to the α-casein fragment, is accommodated in the C-shaped hydrophobic pocket, with a different angle from that of rA-Phe in the complex of LF-transferase with rA-Phe. The α-amino group of Phe and the main-chain amide of Arg in the peptide form hydrogen bonds with the Oε atom of Gln 188, and strong electrostatic interactions occur between the guanidinium group of Arg of the product peptide and the Oε atoms of Glu 156. The surface area of the binding site for the side chain of Arg, adjacent to the hydrophobic region for Phe binding, is highly negative (Fig. 2b). These explain the substrate specificity of LF-transferase, in which proteins bearing N-terminal Arg or Lys can be the substrates as acceptor proteins, and the main recognition element in the acceptor protein by LF-transferase is an N-terminal, basic amino acid residue[17]. The mutation of Glu 156 to either Ala or Gln abolishes the activity for both phenylalanyl and leucyl transfer[19], and the mutation of Glu 156 to Asn severely impairs the transfer activity because the K_m value of α-casein was elevated by a factor of more than ten. The mutation of Glu 156 to Asp reduces the activity, but it still remains at about 21% of that of the wild type (Table 1). These biochemical studies agree well with the recognition of the N-terminal Arg of proteins by Glu 156 in the negatively charged pocket of LF-transferase.

Dynamics of peptide-bond formation

When the two separate LF-transferase complex structures—the complex with rA-Phe and the one with the product peptide bearing an N-terminal Phe (FRYLG)—are superposed, local structural differences between the two structures are apparent. In the product-bound structure, the conformations of Trp 49 and Trp 111 are different from those in the rA-Phe-bound structure (Fig. 3a). The Tyr residue of the product peptide is sterically constrained with the side chain of Trp 111 in the rA-Phe-bound structure. For the accommodation of the peptide, the side chain of Trp 111 flips, which rotates the side chain of Trp 49 (Fig. 3a). The conformation of Gln 188 in the product-bound structure of LF-transferase also differs from that of Gln 188 in the rA-Phe-bound structure (Fig. 3b). An intermolecular hydrogen bond between the Nε atom of Gln 188 and the Oε atom of Glu 156 is formed in the rA-Phe-bound structure. On the other hand, in the product-bound structure, the hydrogen bond breaks and the Oε atom of Glu 156 forms a hydrogen bond with the guanidinium group of Arg of the product peptide, and the Oε atom of Gln 188 forms a hydrogen bond with the main-chain amide of Arg of the product peptide. This conformational change of Gln 188 might be accompanied by the catalysis of peptide-bond formation by LF-transferase, as described below.

The phenylalanyl moiety of rA-Phe is superposed with the Phe of the product peptide in a different orientation (Fig. 3a). The N-terminal Phe of the product peptide is docked in the C-shaped hydrophobic pocket in an inverted manner, in which the benzyl group of Phe is rotated 180 degrees relative to that of rA-Phe. The chemical reaction of peptide-bond formation proceeds by the nucleophilic attack of the α-amino group of the amino acid on the carbonyl carbon of the ester bond of the aminoacyl-tRNA, via a tetrahedral intermediate state. Therefore, the different accommodation of the benzyl groups in the two structures reflects the inversion of the benzyl group after the peptide-bond formation reaction, via the tetrahedral intermediate state.

The adenosine of rA-Phe in the rA-Phe-bound structure is sterically constrained with the amino acid residues RYL of the product-peptide in the superposition. When the LF-transferase crystals were soaked in a solution containing only a peptide bearing an N-terminal Arg, the electron density corresponding to the peptide was barely visible (data not shown). This reflects that Phe or Leu attached to the 3′ terminus of tRNA enters the hydrophobic pocket, before access of the N-terminal side chain of the basic amino acid of the protein to the adjacent, negatively charged pocket. The previous tRNA-docking model, where the tRNA D-stem interacts with LF-transferase, suggested that the disruption or bending of the 3′-acceptor region of the aminoacyl-tRNA is required for the reaction[19,21].

Table 1 | Kinetic parameters of mutant LF-transferase

Variant name	K_m (μm)	α-casein k_{cat} (min^{-1})	Relative (k_{cat}/K_m)
Wild type	5.4	2.6	1.0
E156A	ND		
E156Q	ND		
E156D	8.6	0.85	0.21
E156N	56.0	0.33	0.012
D186A	ND		
Q188A	10.0	0.27	0.056
Q188E	20.0	0.79	0.082
Q188D	33.0	0.36	0.023
Q188N	12.0	2.8	0.48
N191D	4.8	0.13	0.056

ND, not determined.

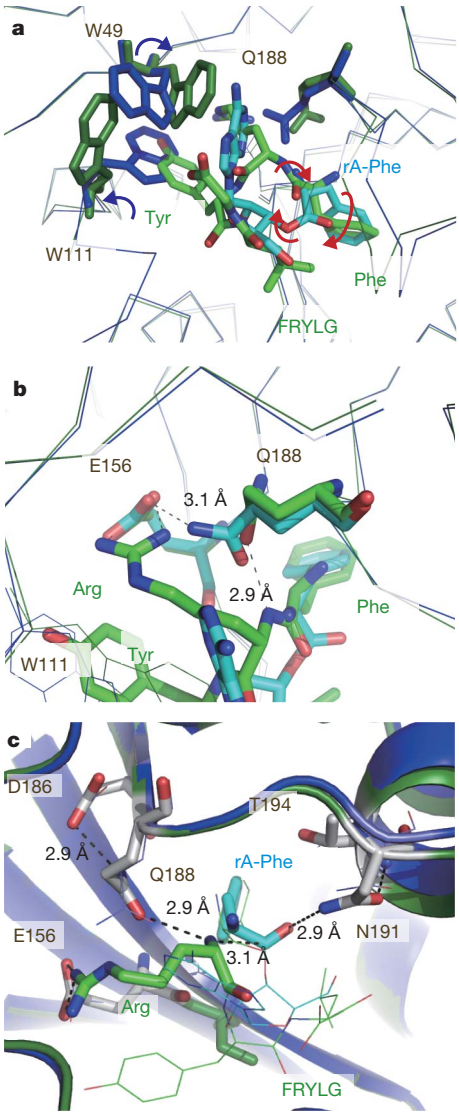


Figure 3 | Superposition of two binary complex structures. **a**, Structural difference between the complex with rA-Phe (blue) and that with the product peptide (green). **b**, Detailed structural difference of the Gln 188 side chains between the two structures. **c**, Superposition of the two complexes. The complex with rA-Phe and that with the product peptide are coloured blue and dark green, respectively. The benzyl group of rA-Phe and the Arg in the product peptide are coloured cyan and green, respectively, and are shown in stick models.

In addition, when phenylalanine or leucine was soaked in the LF-transferase crystals, a clearer electron density of the amino acid was observed (Supplementary Fig. 6), and even rA-Phe can be a substrate (Supplementary Fig. 4). The possible interaction between the D-stem of the aminoacyl-tRNA and LF-transferase could facilitate the positioning of the 3′-terminal region of tRNA in the vicinity of the catalytic pocket of LF-transferase, and the hydrophobic amino acid moiety may be sufficient for aminoacyl-tRNA recognition by LF-transferase. These interactions reflect that no specific hydrogen bond is formed between adenine and LF-transferase (Fig. 1b). Thus, the intrinsic mobile property of the 3′-terminal adenosine of the aminoacyl-tRNA might enable the adenosine to flip on the access of an N-terminal Arg of an acceptor protein to the negatively charged pocket, which is followed by the conformational change of Trp 111, to accommodate the acceptor peptide (Fig. 2b).

Catalytic mechanism

The superposition of the two structures allows the catalytic mechanism of peptide-bond formation by LF-transferase to be deduced (Figs 3c, 4, and Supplementary Fig. 1). The newly synthesized peptide bond (between Phe and Arg) is in the vicinity of the carbonyl carbon of the ester bond of rA-Phe, indicating that the α-amino group of the N-terminal Arg of a protein can attack the carbonyl carbon of the ester bond of the aminoacyl-tRNA. The proximity of the Oε atom of Gln 188 to the peptide bond being synthesized in the product-bound structures suggests that, in the ternary complex, this atom is involved in the catalysis. Presumably, it abstracts or attracts a proton from the α-amino group of the N-terminal Arg, and acts as the general base. The Oε atom of Gln 188 itself cannot act as a general base to abstract a proton from the α-amino group of the N-terminal Arg of proteins. In the binary complex of LF-transferase with the product peptide, the Oδ atom of Asp 186 forms a hydrogen bond with the Nε atom of Gln 188. Therefore, Asp 186 seems to accept a proton from the Nε atom of Gln 188 to help Gln 188 become oriented towards the α-amino group of the N-terminal Arg of the peptide, to act as a general base and to attract a proton from the adjacent α-amino group of the N-terminal Arg. As described, the access of the acceptor protein bearing an N-terminal Arg induces the hydrogen-bond break between the Nε atom of Gln 188 and the Oε atom of Glu 156 (Fig. 3b). This leads to the recognition of the N-terminal Arg of the peptide by the Oε atom of Glu 156, and to the hydrogen-bond formation between the Oδ atom of Asp 186 and the Nε atom of Gln 188. The electron relays between Asp 186 and Gln 188, on binding of the acceptor protein bearing an N-terminal Arg, facilitate the nucleophilic attack of the α-amino group of Arg on the carbonyl carbon of the esterified aminoacyl-tRNAs, and lead to peptide-bond formation. The Nδ atom of Asn 191 is in the proximity of the carbonyl oxygen of the ester bond of the aminoacyl-tRNA analogue, and the Oδ atom of Asn 191 forms a hydrogen bond with the main-chain amide of Thr 194 (Fig. 3b). The hydrogen bond between the Nδ atom of Asn 191 and the carbonyl oxygen of the ester bond may increase the polarity of the carbonyl oxygen of the ester bond, which, in turn, may enhance the electrophilicity of the carbonyl carbon for attack by the nucleophile. In addition, the tetrahedral intermediate of the oxyanion during the peptide-bond formation by LF-transferase may be stabilized by the interaction between the oxyanion and the Nδ atom of Asn 191 (Fig. 3c). Finally, the proton abstracted from the α-amino group by Gln 188 is transferred to the 3′-oxygen of the tRNA, which is liberated with the completion of peptide-bond formation.

The proposed catalytic mechanism of peptide-bond formation by LF-transferase is analogous to the reverse reaction of the acylation step observed in the peptide hydrolysis reaction by serine proteases, such as chymotrypsin[22,23]. In the model for the catalytic mechanism of LF-transferase (Fig. 4), the aminoacyl-tRNA corresponds to the acyl-Ser 195 of chymotrypsin, and Gln 188 and Asp 186 of LF-transferase correspond to His 57 and Asp 102 of chymotrypsin, respectively, which are involved in the electron relay (Supplementary Fig. 7). The role of the Nδ atom of Asn 191 of LF-transferase corresponds to that of the main chain of the amide of Ser 195 and Gly 193 of chymotrypsin, which stabilizes the tetrahedral intermediate. Biochemical studies (Table 1) revealed that the mutation of Gln 188 to Ala severely impaired the activity of LF-transferase[19] (Supplementary Fig. 8), and the mutation of Gln 188 to either Glu or Asp also impaired the activity significantly. However, the mutation of Gln 188 to Asn did not seriously affect the activity, which remained at about 48% of that of the wild type, without affecting the k_{cat}. Moreover, the mutation of Asp 186 to Ala severely impaired the activity of LF-transferase, and the mutation of Asn 191 to Asp significantly affected the activity, with the k_{cat} decreased by a factor of 20, without affecting the K_m value, compared to the wild-type LF-transferase. These biochemical studies agree with the proposed catalytic mechanism of peptide-bond formation by LF-transferase.

Concluding remarks

In this paper, the molecular basis for the non-ribosomal protein-based peptide-bond formation by LF-transferase is presented, as supported by structural and biochemical studies. The mechanism is similar to the reverse of the acylation step of proteolysis by serine proteases, which was originally proposed for the ribosome[24]. However, peptide-bond formation on the ribosome proceeds in a different manner. The ribosome uses the substrate-assisted proton shuttle mechanism for peptide-bond formation, in which the proton shuttle occurs via the 2′-hydroxyl of A76 of the P-site tRNA, and is assisted by neighbouring ribosomal RNA and water molecules. The reaction does not involve chemical catalysis by the proteins, but instead it involves conformational and environmental changes of the active site[25–28]. On the other hand, as presented here, the peptide-bond formation by LF-transferase involves a protein-based chemical reaction, in which the electron relay from Asp 186 to Gln 188 is the underlying mechanism (Supplementary Fig. 9). Therefore, the peptide-bond formations by LF-transferase and by the ribosome proceed in the unique environmental topologies of their active sites and via different catalytic mechanisms.

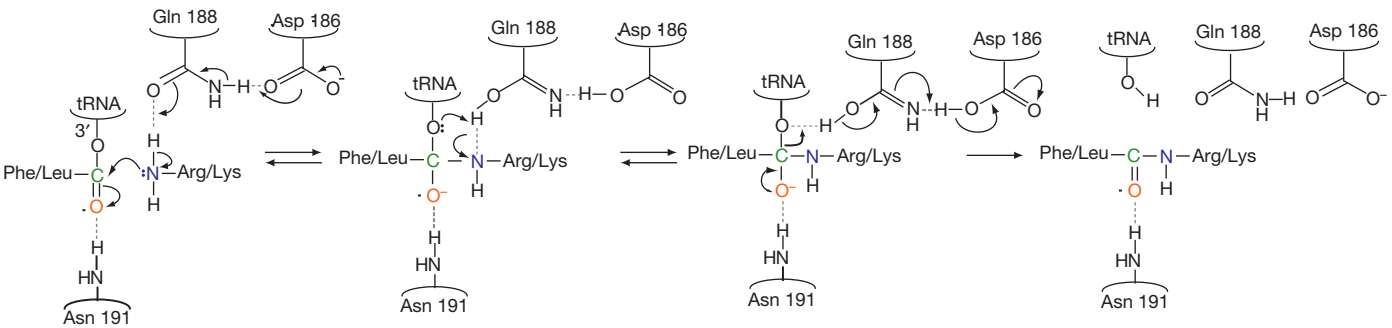


Figure 4 | A model of the catalytic mechanism for peptide-bond formation by LF-transferase.

METHODS SUMMARY

LF-transferase was crystallized as described[19]. Crystal structures were solved by molecular replacement, using the refined *apo*-LF-transferase structure[19] as a search model.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 June; accepted 13 August 2007.
Published online 23 September 2007.

1. Varshavsky, A. The N-end rule. *Cell* **69,** 725–735 (1992).
2. Varshavsky, A. The N-end rule: functions, mysteries, uses. *Proc. Natl Acad. Sci. USA* **93,** 12142–12149 (1996).
3. Mogk, A., Schmidt, R. & Bukau, R. The N-end rule pathway for regulated proteolysis: prokaryotic and eukaryotic strategies. *Trends Cell Biol.* **17,** 165–172 (2007).
4. Rao, H., Uhlmann, F., Nasmyth, K. & Varshavsky, A. Degradation of a cohesin subunit by the N-end rule pathway is essential for chromosome stability. *Nature* **410,** 955–959 (2001).
5. Varshavsky, A. The N-end rule and regulation of apoptosis. *Nature Cell Biol.* **5,** 373–376 (2003).
6. Hu, R. G. *et al.* The N-end rule pathway as a nitric oxide sensor controlling the levels of multiple regulators. *Nature* **437,** 981–986 (2005).
7. Varshavsky, A. The ubiquitin system. *Trends Biochem. Sci.* **22,** 383–387 (1997).
8. Nandi, D., Tahiliani, P., Kumar, A. & Chandu, D. The ubiquitin–proteasome system. *J. Biosci.* **31,** 137–155 (2006).
9. Dougan, D. A., Mogk, A., Zeth, K., Turgay, K. & Bukaku, B. AAA+ proteins and substrate recognition, it all depends on their partner in crime. *FEBS Lett.* **529,** 6–10 (2002).
10. Erbse, A. *et al.* ClpS is an essential component of the N-end rule pathway in *Escherichia coli*. *Nature* **439,** 753–756 (2006).
11. Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. The N-end rule in bacteria. *Science* **254,** 1374–1377 (1991).
12. Hegde, S. S. & Shrader, T. E. FemABX family members are novel nonribosomal peptidyltransferases and important pathogen-specific drug targets. *J. Biol. Chem.* **276,** 6998–7003 (2001).
13. Benson, T. E. *et al.* X-ray crystal structure of *Staphylococcus aureus* FemA. *Structure* **10,** 1107–1115 (2002).
14. Biarrotte-Sorin, S. *et al.* Crystal structures of *Weissella viridescens* FemX and its complex with UDP-MurNAc-pentapeptide: insights into FemABX family substrates recognition. *Structure* **12,** 257–267 (2004).
15. Vetting, M. W. *et al.* Structure and functions of the GNAT superfamily of acetyltransferases. *Arch. Biochem. Biophys.* **433,** 212–226 (2005).
16. Kaji, H., Novelli, G. D. & Kaji, A. A soluble amino acid-incorporating system from rat liver. *Biochim. Biophys. Acta* **76,** 474–477 (1963).
17. Soffer, R. L. Peptide acceptors in the leucine, phenylalanine transfer reaction. *J. Biol. Chem.* **248,** 8424–8428 (1973).
18. Ichetovkin, I. E., Abramochkin, G. & Shrader, T. E. Substrate recognition by the leucyl/phenylalanyl-tRNA-protein transferase. Conservation within the enzyme family and localization to the trypsin-resistant domain. *J. Biol. Chem.* **272,** 33009–33014 (1997).
19. Suto, K. *et al.* Crystal structures of leucyl/phenylalanyl-tRNA-protein transferase and its complex with an aminoacyl-tRNA analog. *EMBO J.* **25,** 5942–5950 (2006).
20. Eriani., G., Delarue, M., Poch, O., Gangloff, J. & Moras, D. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* **347,** 203–206 (1990).
21. Abramochkin, G. & Shrader, T. E. Aminoacyl-tRNA recognition by the leucyl/phenylalanyl-tRNA-protein transferase. *J. Biol. Chem.* **271,** 22901–22907 (1996).
22. Blow, D. M., Briktoft, J. J. & Hartley, B. S. Role of a buried acid group in the mechanism of action of chymotrypsin. *Nature* **221,** 337–340 (1969).
23. Steitz, T. A. & Shulman, R. G. Crystallographic and NMR studies of the serine proteases. *Annu. Rev. Biophys. Bioeng.* **11,** 419–444 (1982).
24. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289,** 920–930 (2000).
25. Beringer, M. & Rodnina, M. V. The ribosomal peptidyl transferase. *Mol. Cell* **26,** 311–321 (2007).
26. Rodnina, M. V., Beringer, M. & Wintermeyer, W. How ribosomes make peptide bonds. *Trends Biochem. Sci.* **32,** 20–26 (2007).
27. Schmeing, T. M., Huang, K. S., Kitchen, D. E., Strobel, S. A. & Steitz, T. A. Structural insights into the roles of water and the 2′ hydroxyl of the P site tRNA in the peptidyl transferase reaction. *Mol. Cell* **20,** 437–448 (2005).
28. Trobro, S. & Åqvist, J. Analysis of predictions for the catalytic mechanism of ribosomal peptidyl transfer. *Biochemistry* **45,** 7049–7056 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Fukai, T. Numata, T. Suzuki and H. Hori for valuable and critical comments, and suggestions for this manuscript. We thank the beam-line staffs at BL-5A, BL-17A and AR-NW 12A of KEK (Tsukuba, Japan) for technical help during data collection, and A. Hamada for technical assistance. This work was supported in part by grants from JSPS, MEXT and the Kurata Memorial Hitachi Science and Technology Foundation (K.T.).

Author Contributions K.T. purified and crystallized the proteins, and K.W., Y.T. and K.T. collected the data and determined the structures. K.S. assisted with the structural analysis, K.W. and K.T. carried out biochemical and mass analyses, Y.S. assisted with the mass analysis, and N.O. and T.W. synthesized analogues. K.W., Y.T. and K.T. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Coordinates and structure factors have been deposited in the Protein Data Bank, under the accession codes 2Z3K, 2Z3L, 2Z3M, 2Z3N, 2Z3O and 2Z3P. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.T. (kozo-tomita@aist.go.jp).

METHODS

Determination of crystal structures. The crystals of LF-transferase were soaked in a solution containing 50 mM Hepes-KOH pH 8.1, 0.7 M tri-sodium tartrate and 5 mM each of either: (1) rA-Phe; (2) dA-Phe; (3) α-casein fragment (RYLGYL); (4) peptide RYLG; (5) rA-Phe and α-casein fragment; (6) rA-Phe and RYLG; (7) peptide FRYLG; or (8) Phe or Leu at 20 °C for 2–7 h. The data were collected at the beam-line AR-NW 12A, BL-5A or BL-17A of KEK (Tsukuba, Japan). All of the data were processed using the program HKL2000[29]. Crystal structures were solved at 2.4–2.85 Å resolutions by molecular replacement with the program MolRep[30]. The model was manually modified using the program O[31], with iterative cycles of refinement with CNS[32] (Supplementary Table 1).

Synthesis of phenyl-tRNAPhe analogue, rA-Phe and dA-Phe. Phenylalanyl adenosine (rA-Phe) and phenylalanyl deoxyadenosine (dA-Phe) were synthesized and purified as previously reported[33,34].

Preparation of aminoacyl-tRNAs. Mutations of the LF-transferase overexpression plasmid were introduced by site-directed mutagenesis, and the mutant enzymes were expressed, purified and quantified as described[19]. The $tRNA^{Phe}$ transcript and *E. coli* Phe-tRNA synthetase (PheRS) were prepared as described[19]. The [^{14}C]-Phe-$tRNA^{Phe}$ transcript was prepared in a solution, containing 100 mM Tris-HCl, pH 7.4, 5 mM $MgCl_2$, 2 mM ATP, 20 mM KCl, 11 μM [^{14}C]-phenylalanine (17 GBq mmol^{-1}; GE Healthcare), 28 μM phenylalanine, 15 μM $tRNA^{Phe}$ transcript and 2 μM PheRS, which was incubated at 37 °C for 30 min. The [^{14}C]-Phe-$tRNA^{Phe}$ was phenol-extracted, precipitated with ethanol and dissolved in 15 mM KOAc, pH 5.3. The amount of [^{14}C]-Phe-$tRNA^{Phe}$ was quantified with a liquid scintillation counter.

***In vitro* assay for aminoacyl-transfer by mutant LF-transferase.** The standard assay for phenylalanine transfer activity was carried out in buffer A (50 mM Tris-Cl, pH 8.0, 100 mM KCl, 10 mM $Mg(OAc)_2$, 1 mM DTT) and 5 μM [^{14}C]-Phe-$tRNA^{Phe}$, 30 μM α-casein (Sigma), and 100 nM purified enzyme. After a 3 min incubation at 37 °C, the reaction mixtures were precipitated with 5% (w/w) TCA, boiled at 90 °C for 30 min and analysed with a liquid scintillation counter. To determine the kinetic parameters for α-casein, the concentrations of the [^{14}C]-Phe-$tRNA^{Phe}$ were fixed at 5 μM, and the α-casein concentration of the reaction mixture was varied (0.5–40 μM). The apparent kinetic parameters K_m and k_{cat} were determined by Lineweaver–Burk plots.

***In vitro* assay for aminoacyl-transfer using 2′-phenylalanyl-$tRNA^{Phe}$ by LF-transferase.** The 2′-phenylalanylated-tRNA with 3′-deoxy-A76 was prepared as described[35]. The reaction mixtures, containing 250 nM LF-transferase, 30 μM α-casein and 2 μM [^{14}C]-phenylalanyl-3′-deoxyA76-tRNA (or [^{14}C]-phenylalanyl-tRNA) in buffer A, were incubated at 37 °C. After 5, 15 and 30 min incubations, the reactions were stopped by adding SDS-loading buffer and incubated at 100 °C for 5 min, to deacylate the aminoacyl-tRNA. The mixtures were separated by 12.5% (v/v) SDS–PAGE, and the gels were dried, visualized and quantified with a BAS-5000 bio-imaging analyser (Fuji Film).

Mass spectrometry analysis of phenylalanine transfer to the α-casein fragment. Phenylalanine transfer reactions were carried out in buffer A, with 200 nM LF-transferase, 10 μM α-casein fragment (RYLGYL; Sigma), and 500 μM rA-Phe (or dA-Phe). After a 1-h incubation at 37 °C, the α-casein fragments in the reaction mixture were purified and desalted by a Nu-Tip C-18 (Glygen), and were eluted with a 50% (v/v) acetonitrile, 0.1% (v/v) TFA solution saturated with the matrix α-cyano-4-hydroxycinnamic acid. Mass measurements were performed using MALDI-TOF (Voyager, Applied Biosystems) in the reflector mode, and were calibrated with des-arg1-bradykinin (*m/z* 904.4), angiotensin I (*m/z* 1,296.7), glu1-fibrinopeptide B (*m/z* 1,570.7), and neurotensin (*m/z* 1,672.9), as standards.

29. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
30. Cowtan, K. An automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM. *Newslett. Protein Crystallogr.* **31,** 34–38 (1994).
31. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
32. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
33. Gottikh, B. P., Krayevsky, A. A., Tarussova, N. B., Purygin, P. P. & Tsilevich, T. L. The general synthetic route to amino acid esters of nucleotides and nucleoside-5′-triphosphates and some properties of these compounds. *Tetrahedron* **26,** 4419–4433 (1970).
34. Chládek, S., Ringer, D. & Žemlička, J. L-Phenylalanine esters of open-chain analog of adenosine as substrates for ribosomal peptidyl transferase. *Biochemistry* **12,** 5135–5138 (1973).
35. Roy, H., Ling, J., Irnov, M. & Ibba, M. Post-transfer editing *in vitro* and *in vivo* by the β subunit of phenylalanyl-tRNA synthetase. *EMBO J.* **23,** 4639–4648 (2004).

LETTERS

A 15.65-solar-mass black hole in an eclipsing binary in the nearby spiral galaxy M 33

Jerome A. Orosz[1], Jeffrey E. McClintock[2], Ramesh Narayan[2], Charles D. Bailyn[3], Joel D. Hartman[2], Lucas Macri[4], Jiefeng Liu[2], Wolfgang Pietsch[5], Ronald A. Remillard[6], Avi Shporer[7] & Tsevi Mazeh[7]

Stellar-mass black holes are found in X-ray-emitting binary systems, where their mass can be determined from the dynamics of their companion stars[1–3]. Models of stellar evolution have difficulty producing black holes in close binaries with masses more than ten times that of the Sun ($>10M_{\odot}$; ref. 4), which is consistent with the fact that the most massive stellar black holes known so far[2,3] all have masses within one standard deviation of $10M_{\odot}$. Here we report a mass of $(15.65 \pm 1.45)M_{\odot}$ for the black hole in the recently discovered system M 33 X-7, which is located in the nearby galaxy Messier 33 (M 33) and is the only known black hole that is in an eclipsing binary[5]. To produce such a massive black hole, the progenitor star must have retained much of its outer envelope until after helium fusion in the core was completed[4]. On the other hand, in order for the black hole to be in its present 3.45-day orbit about its $(70.0 \pm 6.9)M_{\odot}$ companion, there must have been a 'common envelope' phase of evolution in which a significant amount of mass was lost from the system[6]. We find that the common envelope phase could not have occurred in M 33 X-7 unless the amount of mass lost from the progenitor during its evolution was an order of magnitude less than what is usually assumed in evolutionary models of massive stars[7–9].

Optical imaging and spectroscopic observations of M 33 X-7 were obtained in service mode with the 8.2-m Gemini North telescope between 18 August and 16 November 2006. The mean optical spectrum is shown in Fig. 1. The radial velocities derived from the 22 usable spectra show a nearly sinusoidal variation when phased on the orbital period of 3.453014 days determined from the X-ray eclipses[5] (Fig. 2b).

Time series photometry was derived from the Gemini images in the Sloan g′ and r′ filters (Fig. 3). Additional photometric data were obtained in the B, V and I filters using the 3.5-m WIYN telescope during 18–21 August and 15–16 September 2006 (see Supplementary Information). The phased light curves show the characteristic ellipsoidal variations of a tidally distorted star (Fig. 3), which have been reported previously for this source[10,11].

The temperature of the companion star was determined by comparing its averaged spectrum (Fig. 1) to a collection of synthetic spectra derived from the OSTAR2002 grid[12]. As is usually the case, there is a strong correlation in the library of models between effective temperature, $T_{\rm eff}$, and surface gravity, log*g*, and thus various combinations of these parameters result in very similar spectra. Fortunately, the dynamical information strongly constrains the allowed value of log*g* to be between 3.65 and 3.75 at the 3σ level. A good match to the observed spectrum is provided by the model with $T_{\rm eff} = 35{,}000$ K, log$g = 3.75$, a metallicity of 10% of the solar value (representative of star clusters in M 33 at this galactocentric distance[13]), and our measured value of the projected rotational velocity of $V_{\rm rot}\sin i = 250 \pm 7$ km s^{-1} (where i is the orbital inclination angle). The formal error ($\pm 1\sigma$) on the temperature is ± 200 K for log$g =$ 3.75. However, given the possibility of a gravity slightly different from log$g = 3.75$ (the grid spacing of the models is 0.25 dex) and the correlation between $T_{\rm eff}$ and log*g*, we adopt a temperature in the range $34{,}000\ {\rm K} \leq T_{\rm eff} \leq 36{,}000$ K, which corresponds to a spectral type of O7III to O8III (ref. 14).

To compute the radius of this O-type star, we adopt a distance modulus of 24.62 ± 0.05 mag (distance $d = 840 \pm 20$ kpc) to M 33 (see Supplementary Information for details). Using a V-band magnitude $V = 18.9 \pm 0.05$, a V-band extinction of 0.53 ± 0.06 mag (ref. 10), and the bolometric corrections derived from the OSTAR2002 grid, we find a radius of $R_2 = (19.6 \pm 0.9)R_{\odot}$ and a luminosity of $\log(L/L_{\odot}) = 5.72 \pm 0.07$ (here $R_{\odot}$ and $L_{\odot}$ are respectively the solar radius and the solar luminosity).

The duration of the X-ray eclipse Θ (measured in degrees of orbital phase) strongly constrains the available parameter space. Because the size of the compact X-ray source ($\lesssim$1,000 km) is vastly smaller than the secondary star, one might expect the X-ray eclipse profile to be a 'square well', with a flat bottom and very abrupt periods of ingress and egress. However, in M 33 X-7 and other X-ray binaries with

Table 1 | Selected parameters for M 33 X-7

Parameter	Value	Parameter	Value
Θ (degrees)	46 ± 1	M_2 ($M_{\odot}$)	70.0 ± 6.9
$T_{\rm eff}$ (K)	34,000–36,000	$r_{\rm d}$	0.45 ± 0.03
$V_{\rm rot}\sin i$ (km s^{-1})	250 ± 7	e	0.0185 ± 0.0077
R_2 ($R_{\odot}$)	19.6 ± 0.9	ω (degrees)	140 ± 27
$\log L_2$ ($L_{\odot}$)	5.72 ± 0.07	Ω	0.903 ± 0.037
$\Delta\phi$	0.0045 ± 0.0014	f_2	0.777 ± 0.017
i (degrees)	74.6 ± 1.0	a ($R_{\odot}$)	42.4 ± 1.5
K_2 (km s^{-1})	108.9 ± 5.7	M ($M_{\odot}$)	15.65 ± 1.45

The uncertainties correspond to $\pm 1\sigma$. There are three key 'observables' that can be used to determine the mass M of the compact object in an X-ray binary. (1) The radial velocity semiamplitude of the secondary star K_2, along with the orbital period P and eccentricity e, determines the mass function: $f(M) = PK_2^3(1-e^2)^{3/2}/(2\pi G) = M^3 \sin^3 i/(M+M_2)^2$, where M_2 is the mass of the secondary star, i is the orbital inclination angle and G is the gravitational constant. In order to solve for M, we must determine M_2 (or M_2/M) and i, for which we use (2) the rotational velocity of the secondary star $V_{\rm rot}\sin i$, and (3) the amplitude of the ellipsoidal light curve. The two preceding observables depend on i, M_2/M and the Roche-lobe filling f_2, which is the radial fraction of the secondary's Roche equipotential lobe along the line of centres that is occupied by the star. For the determination of Θ, see the Supplementary Information. The measurements of $T_{\rm eff}$ and $V_{\rm rot}\sin i$ were derived directly from the spectra. R_2 and log L_2 were derived from the temperature, apparent magnitude, extinction and the distance. The next nine parameters were determined by fitting the radial velocity curve and light curves simultaneously using the ELC code[15]. The final two parameters are fixed by those given above. See Fig. 3 legend for parameters not defined in the main text.

[1]Department of Astronomy, San Diego State University, 5500 Campanile Drive, San Diego, California 92182-1221, USA. [2]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. [3]Department of Astronomy, Yale University, PO Box 208101, New Haven, Connecticut 06520-8101, USA. [4]National Optical Astronomy Observatory, 950 North Cherry Avenue, Tucson, Arizona 85719, USA. [5]Max-Planck-Institut für extraterrestrische Physik, Giessenbachstraße, D-85741 Garching, Germany. [6]MIT Kavli Institute for Astrophysics and Space Research, 77 Massachusetts Avenue, 37-287, Cambridge, Massachusetts 02139, USA. [7]Wise Observatory, Tel Aviv University, Tel Aviv 69978, Israel.

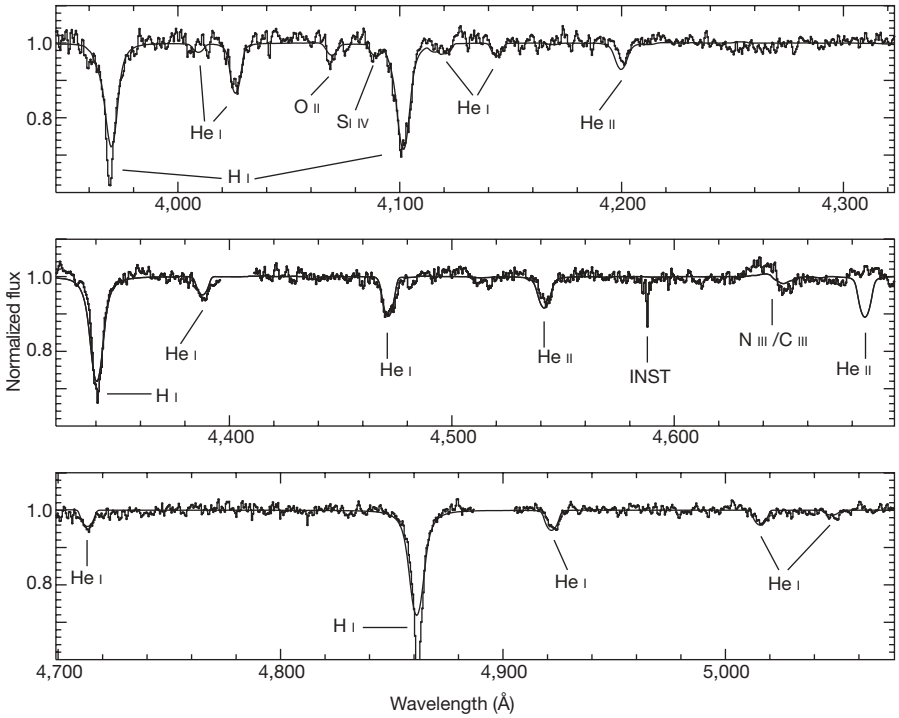


Figure 1 | Mean optical spectrum of M 33 X-7. The spectrum shown here (connected dots), which was extracted with the SPECRES package in IRAF[22], is the sum of the 22 individual spectra that have been velocity-shifted to the rest frame of the secondary star. The solid line is the model spectrum described in the text. The data were obtained using the GMOS instrument on the 8.2-m Gemini-North Telescope with the B1200 grating ($\lambda_c = 4{,}650$ Å) and a 0.5″ slit rotated to a position angle of 215.6°, which was defined by M 33 X-7 and a nearby pair of stars 0.9″ to the southwest (Supplementary Information). Twenty-four 40-min spectra were acquired in service mode between 2006 August 18 and November 16 in good seeing of always <0.8″. The two observations obtained on 2006 September 17 are suspect and will not be considered here. The initial bias subtraction, flat-fielding, and wavelength calibrations were performed using the GMOS package in IRAF. In the two-dimensional spectra, the overlap of the profiles of M 33 X-7 and the nearby pair of stars was modest. The optimal extraction of one-dimensional spectra was done in two ways. (1) Routines in the GMOS package were used with the spectral extraction aperture adjusted so that light from the nearby pair of stars was not included, which resulted in at most about a 20% loss of light from M 33 X-7 (Supplementary Information). The final extracted spectra had signal-to-noise ratios of 20 or more per 0.47 Å pixel near Hβ. Numerous nebular emission lines from the surrounding H II region[23] are seen in these spectra, including the Balmer lines Hβ–Hε, [O III] near 4,363, 4,969 and 5,007 Å, and weak He I lines near 4,026, 4,471, 4,921 and 5,015 Å. The He II line near 4,686 Å and the N III lines near 4,640 Å appear in emission. The quality of the wavelength stability was checked by measuring the radial velocity of the brightest nebular line, [O III] 5,007 Å. Its average heliocentric velocity in the 22 spectra is $-131.2 \pm 1.5\,\mathrm{km\,s^{-1}}$ ($\pm 1\sigma$); for comparison, the velocity of M 33 in the NASA Extragalactic Database is $-179 \pm 3\,\mathrm{km\,s^{-1}}$. (2) Routines in the SPECRES package were used to deblend the spatial profiles of M 33 X-7 and the nearby pair of stars and to remove the nebular lines before optimally extracting one-dimensional spectra. However, the resulting spectra have lower signal-to-noise ratios than the spectra extracted with the GMOS routines (Supplementary Information).

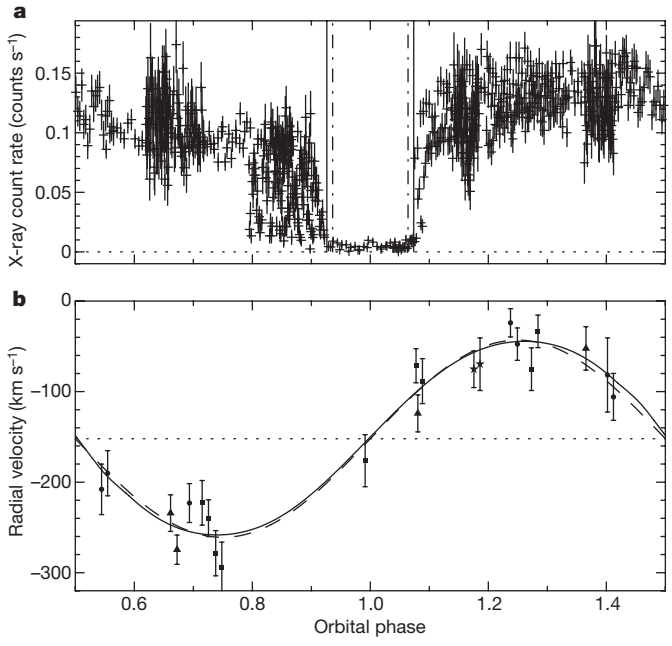


Figure 2 | Phased X-ray light curve and radial velocity curve for M 33 X-7. **a**, The Chandra ACIS light curve in the 0.5–5 keV energy band. Complete orbital phase coverage is achieved here using all 17 available ACIS observations, including five observations (ObsIDs 1730, 6376, 6385, 6387 and 7344) not present in Fig. 1 of ref. 5. The count rates are corrected for vignetting and for the difference between the responses of the ACIS-S and ACIS-I detectors for the single ACIS-I observation (ObsID 6378). The solid vertical lines denote an X-ray eclipse duration of $\Theta = 53°$ (ref. 5), which includes the effects of an extended wind from the O-star. The dash-dotted vertical lines denote an eclipse duration of $\Theta = 46°$, which corresponds solely to an eclipse by the photosphere of the O-star. **b**, The radial velocity curve derived from the Gemini spectra (extracted using the GMOS IRAF package) with the best-fitting model shown as a solid line. The dashed line is the best-fitting sinusoid. The radial velocities were derived by cross-correlating the spectra against a synthetic spectrum (Fig. 1) over the wavelength ranges 4,150–4,300 and 4,521–4,578 Å. These bands include two He II lines, 4,200 Å and 4,541 Å, which are uncontaminated by nebular lines. Radial velocities obtained in 2006 August are denoted by circles, 2006 September by squares, 2006 October by stars, and 2006 November by triangles. Using the orbital period of 3.453014 days determined from the X-ray eclipses[5], a sine fit to the 22 velocities yields $K_2 = 108.9 \pm 6.4\,\mathrm{km\,s^{-1}}$, systemic velocity $\gamma = -152 \pm 5\,\mathrm{km\,s^{-1}}$, and $T_0 = \mathrm{HJD}\ 2{,}453{,}967.157 \pm 0.048$. Here, T_0 is the predicted time of mid-X-ray eclipse, which is in full agreement with that of ref. 5—they differ by 95.001 ± 0.014 orbital cycles. The value of the mass function, which is the absolute minimum mass of the compact object, is $f(M) = (0.46 \pm 0.08)M_\odot$. Error bars in **a** and **b** are $\pm 1\sigma$.

Table 2 | Dynamical measurements of massive stellar black holes

X-ray source	Optical/infrared counterpart	Black hole mass ($M_\odot$)	Secondary star mass ($M_\odot$)
GRS 1915+105	V1487 Aql	14.0 ± 4.4	0.81 ± 0.53
GS 2023+338	V404 Cyg	12 ± 2	0.6
A 0620−00	V616 Mon	10 ± 5	0.6
GS 2000+25	QZ Vul	10 ± 4	0.5
XTE J1550-564	V381 Nor	9.6 ± 1.2	NA
4U 1543−47	IL Lup	9.4 ± 1.1	2.5
Cyg X-1	HDE 226868	>4.8	>11.7
LMC X-1	NA	8–20	NA

Masses for GRS 1915+105 are from ref. 18, and all others are taken from the compilation in ref. 2 and citations therein. The uncertainties correspond to $\pm 1\sigma$. NA, not available.

massive companions, the observed eclipse profile deviates from this idealized picture. The transitions into and out of eclipse are more gradual because of absorption of the X-rays in the stellar wind that thickens near the O-star. The non-zero X-ray intensity in full eclipse results from X-rays that are scattered into our line-of-sight by this wind around the O-star. The erratic X-ray variability before eclipse (Fig. 2a) is presumably caused by absorption in the gas that is streaming from the O-star to the black hole. The period of egress is free of such effects, and we focus our attention there. We identify the eclipse width of $\Theta = 53 \pm 2.2^\circ$ (ref. 5) as the onset of the steep egress feature (solid line in Fig. 2a). In the Supplementary Information, we show that the eclipse width of $\Theta = 53^\circ$ is consistent with absorption in the stellar wind, whereas the true eclipse by the stellar photosphere corresponds to $\Theta = 46 \pm 1^\circ$ (dashed-dotted line in Fig. 2a).

We used a light-curve synthesis code[15] to find the optimal model of the binary system. Figure 3 shows the synthetic light curves for the best-fitting model, which is schematically illustrated in Fig. 4. The best-fitting model parameters and derived astrophysical parameters are summarized in Table 1. The mass of the compact object is $M = (15.65 \pm 1.45)M_\odot$, and the mass of the O-star is $M_2 = (70.0 \pm 6.9)M_\odot$, which puts it among the most massive stars whose masses are well-determined[16]. The effective radii of the Roche lobes are $21.8R_\odot$ and $10.8R_\odot$ for the O-star and black hole, respectively. From evolutionary models of single stars[7–9], the age of the O-star is estimated to be between about 2 and 3 million years. We also note in passing that the O-star is roughly a factor of three less luminous than expected from the evolutionary models.

With $M = (15.65 \pm 1.45)M_\odot$, M 33 X-7 is the most massive stellar black hole known (see Table 2). The mass of V404 Cyg is $(12 \pm 2)M_\odot$ and the masses of the 18 other black holes, save one, are $\lesssim 10M_\odot$, or they are quite imprecise. The one contender is GRS 1915+105, with a mass of $(14.0 \pm 4.4)M_\odot$ (refs 17, 18). However, the 30% precision of the measurement is poor. Furthermore, there are reasons for questioning the reliability of this impressive and pioneering result on a difficult system. For example, the spectroscopic orbital period[17] is 9% longer than the recently-determined and precise photometric period[19]. Furthermore, because of the large X-ray luminosity, the late-type secondary star contributes "only a few per cent of the K-band brightness"[17]; hence the radial velocity curve may be significantly distorted[20]. By comparison, the mass estimate and eclipse ephemeris for M 33 X-7 are exceptionally precise, and X-ray heating is a minor effect.

M 33 X-7 is a key system in the study of high-mass stars, high-mass X-ray binaries, and high-mass black holes. An $\sim 16M_\odot$ black hole paired with an $\sim 70M_\odot$ secondary with a separation of only $\sim 42R_\odot$ is very difficult to explain using stellar evolutionary models. As the radius of the black hole progenitor would have been much larger than the current orbital separation[7–9], the two stars must have been brought closer together by some kind of 'common envelope' phase, which results in a significant amount of mass lost from the progenitor, and very little mass gained by the secondary[6]. On the other hand,

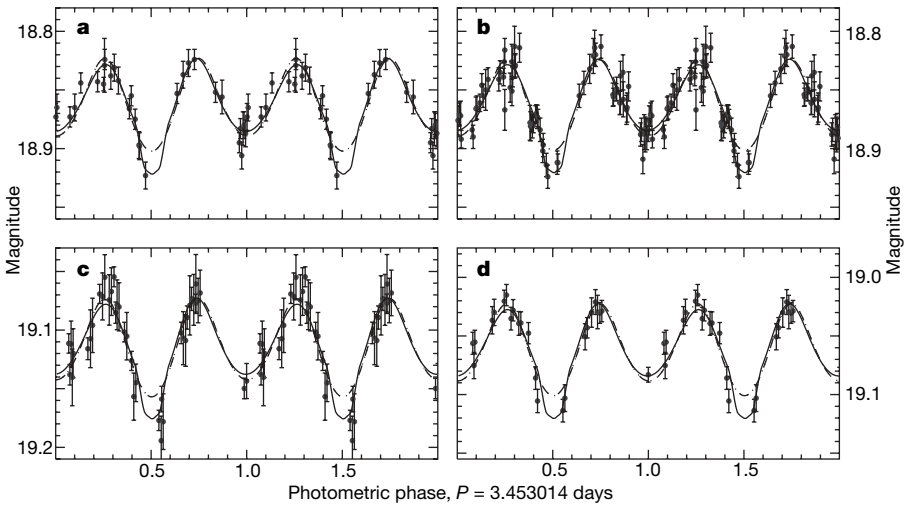


Figure 3 | Optical light curves. **a**, The *B* light curve from ref. 10; **b**, the *V* light curve from ref. 10; **c**, Gemini g′ light curve; **d**, Gemini r′ light curve. The photometric time series were derived using the ISIS image subtraction software[24] together with DAOPHOT[25], which was used to determine the reference flux. The error bars are $\pm 1\sigma$ statistical. The ELC code[15] was used to find the optimal binary model. These light curves and the radial velocities shown in Fig. 2b were used as input data. In addition, we have three other constraints: the radius of the O-star of $R_2 = (19.6 \pm 0.9)R_\odot$, the projected rotational velocity of the O-star of $V_{\rm rot}\sin i = 250 \pm 7\ {\rm km\ s^{-1}}$, and the width of the X-ray eclipse of $\Theta = 46 \pm 1^\circ$ (Supplementary Information). We note that the eclipse duration and known radius are strong constraints that are unavailable for Galactic black hole binaries. In deriving the models, we initially fitted for six parameters: i, K_2, M_2 (see the Table 1 legend), R_2, $T_{\rm eff}$, and a phase shift $\Delta\phi$, which is used to account for small uncertainties in the ephemeris. The first three of these parameters, along with the established orbital period P, determine the scale of the binary, including the dimensions of the Roche equipotential lobes. The value of R_2 then determines the Roche-lobe filling factor f_2. With the geometry of the star fully specified, $T_{\rm eff}$ and the gravity darkening law ($T \propto g^{1/4}$) determine the distribution of temperatures over the surface of the star. No parameterized limb darkening is required because we computed the specific intensities from the OSTAR2002 grid. Likewise, X-ray heating has been accounted for, and is anyway a minor correction ($\Delta T \leq 100$ K) because of the star's extreme luminosity. After several initial trial runs, we found that the fits were improved by (1) adding a faint accretion disk around the compact object with a fractional radius $r_{\rm d}$, (2) allowing the orbit to be slightly non-circular (adds eccentricity e and argument of periastron ω as free parameters), and (3) allowing the O-star to rotate slightly non-synchronously with the orbit, which is parameterized by $\Omega = P_{\rm orb}/P_{\rm rot}$ (we assume the star's rotation axis is perpendicular to the orbital plane). The solid lines show the best-fitting model, and the dash-dotted lines show the best-fitting models with a circular orbit and no accretion disk. The genetic optimizer code was run five times with different initial random parameter sets and the grid search optimizer was run many hundreds of times to refine the solution and define confidence limits on the fitted and derived parameters (Supplementary Information).

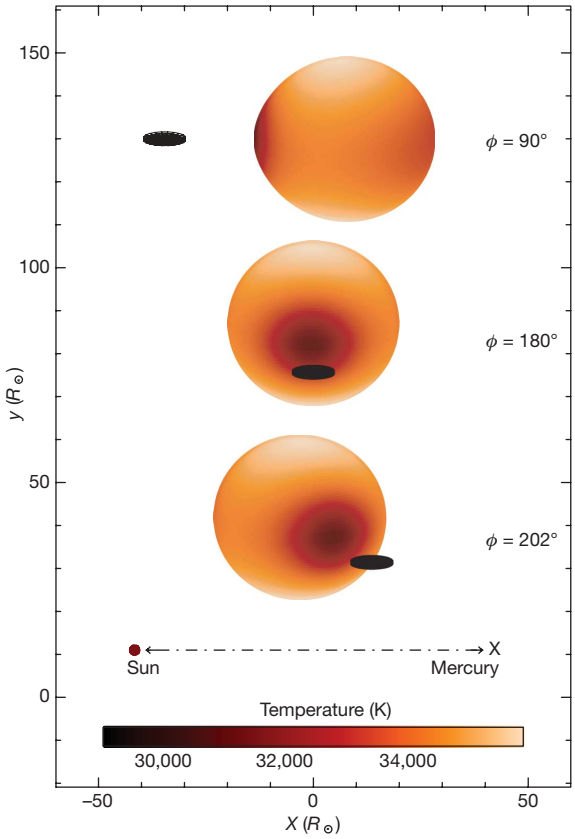


Figure 4 | Schematic diagram of M 33 X-7. The companion star and the accretion disk surrounding the black hole are shown to scale, as seen projected onto the plane of the sky at three orbital phases. The colours on the star represent temperatures (not intensities), with cooler temperatures shown by darker colours, as denoted on the bar. The distance between the Sun and Mercury is indicated, and the figure is scaled in solar radii.

in order for the core mass to remain large enough to produce an $\sim 16 M_{\odot}$ black hole, the outer envelope of the progenitor needs to be intact until core helium burning is completed[4]. Hence we require the common envelope phase to begin only after core helium burning in the progenitor is complete (case C mass transfer[6]). There are two requirements for a common envelope phase to start during case C mass transfer. First, the mass donor needs to be at least 1.2 times more massive than the secondary at the start of mass transfer[21]. Second, the radius of the mass donor at the end of core helium burning needs to be larger than its radius at the end of core hydrogen burning. If the second condition is not met, the common envelope phase begins before core helium burning is complete, and the stripped core loses much of its mass via strong winds in its subsequent evolution and thus cannot make a massive black hole[4].

Assuming no large change in the present-day mass loss rate of $2.6 \times 10^{-6} M_{\odot}\ \mathrm{yr}^{-1}$ (see Supplementary Information), the secondary star has lost between about $5.2 M_{\odot}$ and $7.8 M_{\odot}$, thereby putting its initial mass near $\sim 80 M_{\odot}$. For a common envelope phase to occur, the progenitor star should have been more massive than $\sim 80 \times 1.2 = 96 M_{\odot}$, which is problematic. According to evolutionary models[7–9], massive stars lose much of their initial mass via winds, and the mass loss rate generally increases with increasing initial mass. For example, even in the extreme case of an initial mass of $120 M_{\odot}$ and a metallicity of 20% solar, the mass after hydrogen burning is $\sim 52.9 M_{\odot}$ and after helium burning is $\sim 17.2 M_{\odot}$ (ref. 9). Furthermore, owing to the large amount of mass loss, the radius of the star after core helium burning is smaller than the radius after core hydrogen burning. For these reasons, a common envelope phase during case C mass transfer seems very unlikely. It would appear that the progenitor star of M 33 X-7 lost roughly an order of magnitude less mass before the common envelope phase ensued than is predicted by the evolutionary models. Finally, we note that there is an additional complication: even if a common envelope is formed, the most likely outcome would be a merger, as the envelopes of massive stars are tightly bound[21]. However, the most massive star considered in ref. 21 was $50 M_{\odot}$, so the detailed computations should be extended to the higher masses relevant for M 33 X-7.

The determination of an accurate mass for M 33 X-7—located at a distance of more than 16 times that of any other confirmed stellar black hole—marks a major advance in our capability to study stellar black holes in Local Group galaxies beyond the Milky Way.

Received 27 June; accepted 29 August 2007.

1. Remillard, R. A. & McClintock, J. E. X-ray properties of black-hole binaries. *Annu. Rev. Astron. Astrophys.* **44,** 49–92 (2006).
2. Charles, P. A. & Coe, M. J. in *Compact Stellar X-ray Sources* (eds Lewin, W. H. G. & van der Klis, M.) 215–265 (Cambridge Univ. Press, Cambridge, UK, 2006).
3. Orosz, J. A. in *A Massive Star Odyssey: From Main Sequence to Supernova* (eds van der Hucht, K. A., Herrero, A. & Esteban, C.) 365–371 (Proc. IAU Symp. 212, ASP, San Francisco, 2003).
4. Brown, G. *et al.* Formation of high mass X-ray black hole binaries. *New Astron.* **6,** 457–470 (2001).
5. Pietsch, W. *et al.* M33 X-7: ChASeM 33 reveals the first eclipsing black hole X-ray binary. *Astrophys. J.* **646,** 420–428 (2006).
6. Tauris, T. M. & van den Heuvel, E. P. J. in *Compact Stellar X-ray Sources* (eds Lewin, W. H. G. & van der Klis, M.) 623–665 (Cambridge Univ. Press, Cambridge, UK, 2006).
7. Schaller, G., Schaerer, D., Meynet, G. & Maeder, A. New grids of stellar models from 0.8 to 120 $M_{\odot}$ at $Z = 0.020$ and $Z = 0.001$. *Astron. Astrophys.* **96** (Suppl.), 269–331 (1992).
8. Meynet, G., Maeder, A., Schaller, G., Schaerer, D. & Charbonnel, C. Grids of massive stars with high mass loss rates. V. From 12 to 120 $M_{\odot}$ at Z=0.001, 0.004, 0.008, 0.020 and 0.040. *Astron. Astrophys.* **103** (Suppl.), 97–105 (1994).
9. Vázquez, G. A., Leitherer, C., Schaerer, D., Meynet, G. & Maeder, A. Models for massive stellar populations with rotation. *Astrophys. J.* **663,** 995–1020 (2007).
10. Pietsch, W. *et al.* The eclipsing massive X-ray binary M33 X-7: New X-ray observations and optical identification. *Astron. Astrophys.* **413,** 879–887 (2004).
11. Shporer, A., Hartman, J., Mazeh, T. & Pietsch, W. Photometric analysis of the optical counterpart of the black hole HMXB M33 X-7. *Astron. Astrophys.* **462,** 1091–1095 (2007).
12. Lanz, T. & Hubeny, I. A Grid of non-LTE line-blanketed model atmospheres of O-type stars. *Astrophys. J.* **146** (Suppl.), 417–441 (2003).
13. Ma, J. *et al.* Spectral energy distributions, ages, and metallicities of star clusters in M 33. *Astron. J.* **122,** 1796–1806 (2001).
14. Heap, S. R., Lanz, T. & Hubeny, I. Fundamental properties of O-type stars. *Astrophys. J.* **638,** 409–432 (2007).
15. Orosz, J. A. & Hauschildt, P. H. The use of the NextGen model atmospheres for cool giants in a light curve synthesis code. *Astron. Astrophys.* **364,** 265–281 (2000).
16. Geis, D. R. in *A Massive Star Odyssey: From Main Sequence to Supernova* (eds van der Hucht, K. A., Herrero, A. & Esteban, C.) 91–100 (Proc. IAU Symp. 212, ASP, San Francisco, 2003).
17. Greiner, J., Cuby, J. G. & McCaughrean, M. J. An unusually massive stellar black hole in the Galaxy. *Nature* **414,** 522–525 (2001).
18. Harlaftis, E. T. & Greiner, J. The rotational broadening and the mass of the donor star of GRS 1915+105. *Astron. Astrophys.* **414,** L13–L16 (2004).
19. Neil, E. T., Bailyn, C. D. & Cobb, B. E. Infrared monitoring of the microquasar GRS 1915+105: detection of orbital and superhump signatures. *Astrophys. J.* **657,** 409–414 (2007).
20. Reynolds, A. P. *et al.* A new mass estimate for Hercules X-1. *Mon. Not. R. Astron. Soc.* **288,** 43–52 (1997).
21. Podsiadlowski, Ph., Rappaport, S. & Han, Z. On the formation and evolution of black hole binaries. *Mon. Not. R. Astron. Soc.* **341,** 385–404 (2003).
22. Lucy, L. B. & Walsh, J. R. Iterative techniques for the decomposition of long-slit spectra. *Astron. J.* **125,** 2266–2275 (2003).
23. Humphreys, R. M. & Sandage, A. On the stellar content and structure of the spiral galaxy M 33. *Astrophys. J.* **44** (Suppl.), 319–381 (1980).
24. Alard, C. Image subtraction using a space-varying kernel. *Astron. Astrophys.* **144** (Suppl.), 363–370 (2000).
25. Stetson, P. B. DAOPHOT — A computer program for crowded-field stellar photometry. *Publ. Astron. Soc. Pacif.* **99,** 191–222 (1987).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Walsh for help with the SPECRES software, I. Hubeny for the use of his model atmosphere codes, and T. Matheson for support with the Gemini Observations. C.D.B. acknowledges support from the US National Science Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.A.O. (orosz@sciences.sdsu.edu).

LETTERS

Nature of the superconductor–insulator transition in disordered superconductors

Yonatan Dubi[1], Yigal Meir[1,2] & Yshai Avishai[1,2]

The interplay of superconductivity and disorder has intrigued scientists for several decades. Disorder is expected to enhance the electrical resistance of a system, whereas superconductivity is associated with a zero-resistance state. Although superconductivity has been predicted to persist even in the presence of disorder[1], experiments performed on thin films have demonstrated a transition from a superconducting to an insulating state with increasing disorder or magnetic field[2]. The nature of this transition is still under debate, and the subject has become even more relevant with the realization that high-transition-temperature (high-T_c) superconductors are intrinsically disordered[3–5]. Here we present numerical simulations of the superconductor–insulator transition in two-dimensional disordered superconductors, starting from a microscopic description that includes thermal phase fluctuations. We demonstrate explicitly that disorder leads to the formation of islands where the superconducting order is high. For weak disorder, or high electron density, increasing the magnetic field results in the eventual vanishing of the amplitude of the superconducting order parameter, thereby forming an insulating state. On the other hand, at lower electron densities or higher disorder, increasing the magnetic field suppresses the correlations between the phases of the superconducting order parameter in different islands, giving rise to a different type of superconductor–insulator transition. One of the important predictions of this work is that in the regime of high disorder, there are still superconducting islands in the sample, even on the insulating side of the transition. This result, which is consistent with experiments[6,7], explains the recently observed huge magneto-resistance peak in disordered thin films[8–10] and may be relevant to the observation of 'the pseudogap phenomenon' in underdoped high-T_c superconductors[11,12].

Superconductivity—the occurrence of zero-resistance state—has been a central issue in solid-state physics for nearly a century. About fifty years after its discovery Bardeen, Cooper and Schreiffer[13] (BCS) explained its microscopic foundation. BCS theory attributes superconductivity to pairing of electrons (Cooper pairs), thus creating a many-body coherent macroscopic wavefunction. Electron pairing defines a global order parameter Δ, characterized by its amplitude and phase. According to BCS theory, the suppression of Δ to zero by increasing temperature T or magnetic field B destroys the superconducting state.

Soon after the emergence of BCS theory, Anderson[1] showed that weak disorder cannot lead to the destruction of pair correlations. Later, Lee and Ma[14] argued that strong disorder gives rise to spatial fluctuations of Δ along with its suppression in comparison with its value for the clean system, leading eventually to the destruction of the superconducting state. Such a superconductor–insulator transition (SIT) has indeed been observed in disordered thin superconducting films[2]. A similar, magnetic-field-driven SIT has also been observed. This transition has provoked vast interest, and phenomenological theories, valid near the transition, have been put forward[15,16].

Here, on the basis of the first numerical investigation of the SIT starting from a purely microscopic model, the following physical scenario has emerged. In the presence of disorder, the local superconducting order parameter $\Delta(\mathbf{r})$ develops strong spatial fluctuations[14,17,18], such that regions of space where the amplitude of Δ is large (called 'superconducting islands') are surrounded by regions with relatively small Δ. The system behaves as a bulk superconductor as long as Δ is different from zero, and the phases of $\Delta(\mathbf{r})$ on two sides of the sample are correlated. Such correlations are established by coherent tunnelling of Cooper pairs between the islands. For weak disorder, increasing T or B suppresses Δ in the entire sample, before the system loses phase rigidity, and thus superconductivity is destroyed in a way similar to BCS theory. On the other hand, for stronger disorder, increasing T or B leads to the breakdown of phase-coherent paths between the edges of the sample, thereby driving a transition to an insulating state, even when the superconducting order parameter is still finite. The persistence of superconducting correlations in the insulating phase should have far-reaching observable physical consequences.

Our starting point is the microscopic two-dimensional disordered negative-U Hubbard model (see Methods). This model describes electrons propagating on a two-dimensional disordered square lattice, subject to mutual attraction when two electrons, with opposite spin projections, occupy the same site. The model is known to generate a superconducting ground state when no disorder is present. We first demonstrate the formation and evolution of superconducting islands by solving the Bogoliubov–de Gennes[19] equations (described in the Methods) in the presence of both disorder and magnetic field. A topographic colour plot of the spatial distribution of $|\Delta(\mathbf{r})|$, the amplitude of Δ, for a given disorder realization and a finite B is shown in Fig. 1a. The fluctuations in $|\Delta|$ are clearly visible, and one can resolve regions of high $|\Delta|$ surrounded by regions of low $|\Delta|$. However, the Bogoliubov–de Gennes mean-field approach neglects phase fluctuations altogether, and all regions with non-vanishing Δ are thus phase-correlated. Consequently, within this approximation, as long as $\langle|\Delta|\rangle$—the spatially averaged $|\Delta|$—fails to vanish, the system behaves as a bulk superconductor. With increasing magnetic field, disorder or temperature, there will be a critical point where $\langle|\Delta|\rangle$ vanishes, and the system loses its superconducting nature. Although such a BCS transition is indeed applicable for weakly disordered systems, we show below that this description breaks down for higher disorder, where phase fluctuations play a crucial role.

To take into account phase fluctuations, here we use a newly developed method[12,20] (see Methods). While neglecting quantum fluctuations, the method allows calculation of thermal averages of phase correlations, thus going beyond the lowest-energy, saddle-point

[1]Department of Physics, Ben Gurion University, [2]The Ilse Katz Center for Meso- and Nano-Scale Science and Technology, Ben Gurion University, Beer Sheva 84105, Israel.

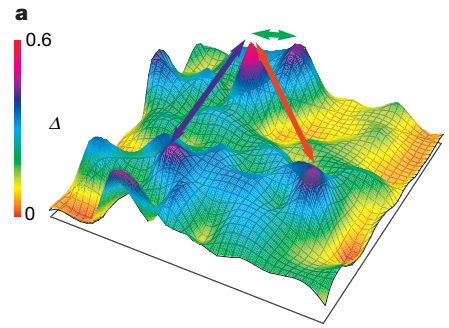


Figure 1 | Spatial fluctuations of the order parameter amplitude and corresponding phase correlations. a, Spatial distribution of $|\Delta|$ for perpendicular field $\phi/\phi_0 = 2.6$ and temperature $T = 0.008t$. The system is of size 12×12 and has an average electron density $\langle n \rangle = 0.92$, with disorder strength $W/t = 1$. Arrows indicate pairs of points in the sample between which the phase correlations were calculated, shown in **b**. With increasing magnetic field the system separates into islands, the phase correlations between them are suppressed (red and blue arrows and stars and triangles), while for points on the same island (green arrow and diamonds) the phases remain correlated.

Bogoliubov–de Gennes solution. In Fig. 1b we plot the magnetic-field dependence of the thermally averaged phase correlations $\langle \cos(\delta\theta_i - \delta\theta_j) \rangle$, where $\delta\theta_i$ is the change of phase of $\Delta(\mathbf{r}_i)$ from its mean-field value, and $\mathbf{r}_i$ and $\mathbf{r}_j$ are different points in the sample, indicated by arrows in Fig. 1a. For the points connected by the green arrow in Fig. 1a, the phase correlations hardly change with B (green curve in Fig. 1b), indicating that these points belong to a coherent superconducting island. However, the points connected by blue and red arrows in Fig. 1a lose their phase coherence with increasing B. Thus, at this field the coherent macroscopic superconducting system separates into phase-uncorrelated superconducting islands.

Using the same method we demonstrate the emergence of a magnetic-field-driven SIT. In Fig. 2 we plot the spatial average of $|\Delta(\mathbf{r})|$ (blue triangles) and the phase correlations (red squares) between the two edges of a superconducting film as a function of B. For weak disorder near half filling (Fig. 2a), the superconducting order parameter vanishes at a critical field. Phase correlations between the two sides of the sample persist until that field is reached. On the other hand, at higher disorder (Fig. 2b) or at lower electron density (which corresponds to effective high disorder, inset of Fig. 2a), the critical field B_c is determined by the loss of phase correlations. The amplitude of the order parameter exhibits no particular feature at the transition, and vanishes at a much higher field. Hence, the nature of this transition is entirely distinct from that at low (or no) disorder (and is probably related to the disordered X–Y model[15]). Above B_c the system displays insulating behaviour, but nevertheless supports superconducting correlations, as long as B is lower than the BCS critical field.

Suppression of phase coherence between the superconducting islands with increasing B is displayed in Fig. 3a–c, where on top of the spatial distribution of $|\Delta(\mathbf{r})|$ we depict the phase correlation of each point on the lattice with the three points of highest $|\Delta|$—the same points as in Fig. 1a. Each colour—red, green, blue—indicates correlation with a different point, so that black (mixture of red, green and blue) corresponds to correlation with all points, and white indicates correlations with none. For zero B, most points are phase-correlated, but as B increases the islands begin to disconnect, eventually becoming well separated. At such fields the system behaves as an insulator, but both unpaired electrons and Cooper pairs coexist and contribute to the transport process. The persistence of pair correlations beyond the SIT accounts for additional experimental findings, such as local superconducting behaviour on the insulating part of the transition[4–7], and the huge magneto-resistance peak observed in these systems[7–9], which was explained by the competition between contributions of Cooper pairs and unpaired electrons[21].

Local measurements of phase correlations are highly daunting, so we propose that the position of the islands and their extent may be experimentally detected by inspecting the dependence of the amplitude of $\Delta(\mathbf{r})$ on a parallel magnetic field $h_{||}$ that couples only to the electron spin. For clean systems, it is well-known[22,23] that such a field leads to an abrupt vanishing of Δ and the destruction of the superconducting state into a spin-polarized state, when the gain in Zeeman energy overcomes the superconducting gap. By solving the Bogoliubov–de Gennes equations in the presence of such parallel field, we verify that in the absence of a perpendicular field (when all phases are correlated), the superconducting gap is indeed destroyed abruptly (purple curve in Fig. 3d). However, for higher perpendicular field (thus decreasing correlations between the phases

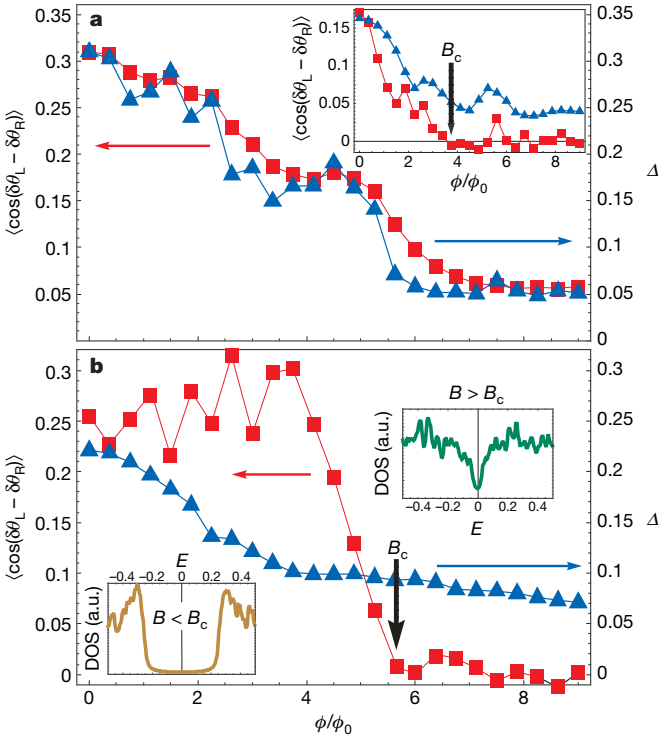


Figure 2 | The superconductor–insulator phase transition with amplitude vanishing and loss of phase coherence. a, Superconducting order parameter amplitude $|\Delta|$ (blue triangles) and phase correlations between the edges ($\theta_{L(\text{or R})}$ stands for order-parameter phases on sites which lie on the left (or right) edge of the sample) of a sample of size 15×5 (red squares), as a function of magnetic field for a weakly disordered sample ($W/t = 0.1$) at electron density $\langle n \rangle = 0.92$ and temperature $T/t = 0.04$. Both $|\Delta|$ and the phase correlations vanish at the same B. **b**, The same for a system with stronger disorder ($W/t = 1$), or lower density, $\langle n \rangle = 0.42$ (inset of **a**). Here the phase correlations vanish long before the amplitude. The insets in **b** show the density of states (DOS) at zero field (brown) and on the insulating side of the transition (green), displaying a pseudo-gap feature similar to that observed in high-T_c superconductors.

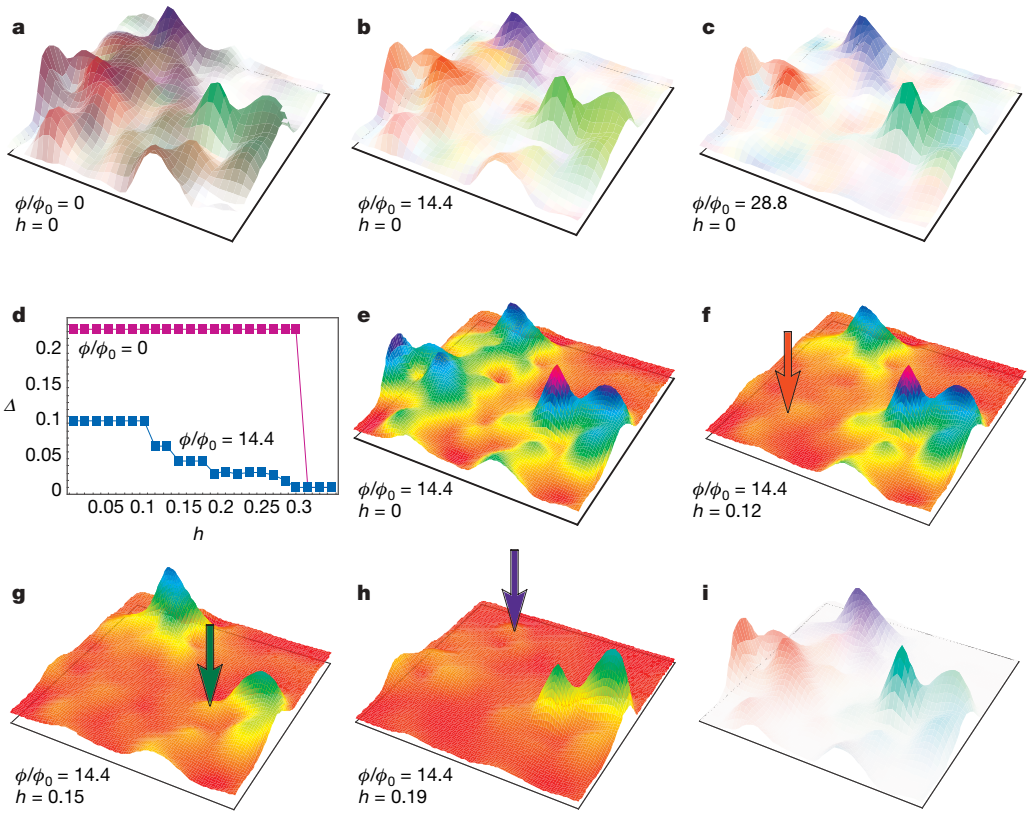


Figure 3 | Superconducting islands observed by phase correlations and by application of a parallel field. **a**–**c**, A spatial map of the phase-correlations: the red, green and blue components of the colour of each point in the sample is proportional to the magnitude of its phase correlations with the three peaks of maximal amplitude (Fig. 1), for different perpendicular magnetic fields, displayed on top of the spatial distribution of the $|\Delta|$. At zero field, phase correlations are long-ranged, but as the magnetic field is increased the system separates into islands with no inter-island correlations. **d**, $\langle|\Delta|\rangle$ as a function of parallel field, for two different values of the perpendicular magnetic field. At zero perpendicular field the superconducting amplitude vanishes abruptly, while at a finite perpendicular field, $\langle|\Delta|\rangle$ decreases in a series of steps, each island at a time. This is demonstrated in **e**–**h**, depicting spatial distribution of $|\Delta|$ for different values of the parallel field. Arrows indicate the position of the superconducting islands, whose $|\Delta|$ vanished at that particular field. **i**, The same distribution of $|\Delta|$, where now each point is coloured by the value of the field at which the amplitude of Δ at that point was suppressed (see **e**–**h**). Comparing with **c** demonstrates that the islands defined this way are identical to those defined by the loss of phase correlations.

of the superconducting islands) we find that $|\Delta|$ vanishes in a step-like manner with $h_{||}$ (blue curve in Fig. 3d), and each step corresponds to the destruction of a different superconducting island. This is depicted in Fig. 3e–h, where the spatial distribution of $|\Delta|$ is plotted for different values of $h_{||}$. The arrows indicate spatial regions where the superconducting gap vanishes at that field. In Fig. 3i we re-plot the amplitude map, in which each point is now coloured according to the field $h_{||}$ at which the local $|\Delta(\mathbf{r})|$ has changed. Comparison with Fig. 3c shows that these regions indeed correspond to the superconducting islands as defined by phase correlations and thus they are directly amenable to local experimental probes.

In our model, we have considered only thermal phase fluctuations (owing to computational constraints). That SIT may be explained in terms of thermal fluctuations accounts for many experimental observations in which the universality of a quantum phase transition is not observed, such as the lack of a universal resistance at the transition[24], temperature dependence of the crossing point[25] and the observed classical X–Y critical exponent[26], and even for the percolation-like behaviour found in some experiments[27–29]. However, it may well be that a similar loss of phase correlations will be driven by quantum fluctuations at low enough temperatures. In fact, recent experiments[7,25] that have explored the competition between thermal and quantum fluctuations (for example, by looking at the dependence on temperature of the crossing point in the resistance–magnetic field plane) demonstrate a continuous crossover from a thermal-fluctuations-driven transition at high temperatures to a quantum-fluctuations-driven transition at low temperatures, the phenomenology of the transition in the two regimes being almost indistinguishable. The non-universality of the critical resistance at the transition may be due to the fact that the dirty boson model for the quantum phase transition does not include the contribution of the unpaired fermions, rather than indicating the irrelevance of the quantum phase-transition scenario.

Finally, while the calculations described above were performed for s-wave superconductors, phase fluctuations have been suggested to be relevant also for high-T_c superconductors[30]. A similar method was recently used[12] to study the phase diagram of a phenomenological model for high-T_c superconductors. The authors[12] found that phase fluctuations can account for several features of the high-T_c superconductors, among them the existence of a disorder-driven pseudo-gap state[11]. To demonstrate the possible relevance of our work, in the inset of Fig. 2b we plot the density of states of the system below (brown) and well above (green) the SIT. Below the SIT the density of states exhibits regular BCS-like superconducting behaviour, while above the transition the density of states exhibits a pseudo-gap feature due to the contribution of the superconducting correlations on the insulating side. For weaker disorder, this feature is only observed at lower density, which might correspond to the fact that the pseudo-gap is solely a feature of underdoped systems. We believe that incorporating phase fluctuations into a microscopic model for the high-T_c superconductors will prove useful in explaining many of the experimental features of these systems.

METHODS SUMMARY

The model. The negative-U Hubbard model is described by the hamiltonian:

$$H=\sum_{i,\sigma}(\varepsilon_i+h_{||}\,\sigma)\,C^{+}_{i\sigma}C_{i\sigma}-t\sum_{<ij>,\sigma}\left(e^{i\phi_{ij}}C^{+}_{i\sigma}C_{j\sigma}+e^{-i\phi_{ij}}C^{+}_{j\sigma}C_{i\sigma}\right)-U\sum_{i}C^{+}_{i\uparrow}C_{i\uparrow}C^{+}_{i\downarrow}C_{i\downarrow} \quad (1)$$

where $C_{i\sigma}$ and $C^{+}_{i\sigma}$ destroy and create an electron with spin σ at site i, respectively. The first term describes the random potential on the two-dimensional lattice, with a possible Zeeman field $h_{||}$, while the second one describes the hopping between nearest-neighbour sites. The phases ϕ_{ij} account for the orbital effects of the magnetic field. The last term describes the attractive interaction between electrons on the same site and is responsible for the emergence of superconductivity. All energies are expressed in units of t, the hopping matrix element. The system is characterized by the relative strength of the attractive interaction U (taken to be $U=2$ throughout the calculation), the disorder parameter W, the parallel magnetic field $h_{||}$, the average electron density n and the perpendicular magnetic field B. W is the range of fluctuations in the on-site energies ε_i whereas B is characterized by the magnetic flux penetrating the sample in units of the quantum flux $\phi_0=hc/e$, where h is Planck's constant, c is the speed of light and e is the charge of the electron).

The partition function for this model is given by:

$$Z=\iint D\{C_i,C_i^{+}\}\exp\left(-\int_0^{\beta}\mathrm{d}\tau\left[\sum_{i\sigma}C^{+}_{i\sigma}(\tau)(-\partial_\tau+\varepsilon_i+h_{||}\sigma)C_{i\sigma}(\tau)-\sum_{<ij>\sigma}(t_{ij}C^{+}_{i\sigma}(\tau)C_{j\sigma}(\tau)+c.c.)-U\sum_i C^{+}_{i\uparrow}(\tau)C^{+}_{i\downarrow}(\tau)C_{i\downarrow}(\tau)C_{i\uparrow}(\tau)\right]\right) \quad (2)$$

where $\beta=1/k_BT$, with k_B the Boltzman constant. *c.c.* denotes complex conjugate and Tr denotes a trace over all possible states. Applying a Hubbard–Stratonovic transformation, with Δ_i the local Hubbard–Stratonovic field, with amplitude $|\Delta_i|$ and phase θ_i, the partition function becomes:

$$Z=\iint D\{\Delta_i,\theta_i\}D\{C_i,C_i^{+}\}\exp\left(-\int_0^{\beta}\mathrm{d}\tau\left[\sum_{i\sigma}C^{+}_{i\sigma}(\tau)(-\partial_\tau+\varepsilon_i+h_{||}\sigma)C_{i\sigma}(\tau)-\sum_{<ij>\sigma}(t_{ij}C^{+}_{i\sigma}(\tau)C_{j\sigma}(\tau)+c.c.)-\sum_i(\Delta_i(\tau)e^{-i\theta_i(\tau)}C^{+}_{i\uparrow}(\tau)C^{+}_{i\downarrow}(\tau)+c.c.)+\sum_i\frac{|\Delta_i(\tau)|^2}{U}\right]\right) \quad (3)$$

The Bogoliubov–de Gennes approximation. The partition function can be evaluated in the saddle-point approximation. Then the effective hamiltonian becomes:

$$H_{\mathrm{BdG}}=\sum_{i,\sigma}(\varepsilon_i+h\,\sigma)\,C^{+}_{i\sigma}C_{i\sigma}-t\sum_{<ij>,\sigma}\left(e^{i\phi_{ij}}C^{+}_{i\sigma}C_{j\sigma}+e^{-i\phi_{ij}}C^{+}_{j\sigma}C_{i\sigma}\right)+\sum_i\left(\Delta_iC^{+}_{i\uparrow}C^{+}_{i\downarrow}+\Delta_i^{*}C_{i\downarrow}C_{i\uparrow}\right) \quad (4)$$

where Δ_i are now constants that obey the self-consistent relations $\Delta_i=-U<C_{i\uparrow}C_{i\downarrow}>$.

H_{BdG} is diagonalized via a Bogoliubov transformation $\gamma_{n\sigma}=\sum\left(u_n(r_i)C^{+}_{i\sigma}+\sigma v_n(r_i)C_{i\bar{\sigma}}\right)$. This yields an equation for the local order parameter Δ_i in terms of the Bogoliubov amplitudes $u_n(i)$ and $v_n(i)$:

$$\Delta_i=|U|\sum_n u_n(i)v_n^{*}(i) \quad (5)$$

$u_n(i)$ and $v_n(i)$ are determined from the Bogoliubov–de Gennes equations[19]:

$$\begin{pmatrix}\hat{\xi} & \Delta_i\\ \Delta_i^{*} & -\hat{\xi}\end{pmatrix}\begin{pmatrix}u_n(i)\\ v_n(i)\end{pmatrix}=E_n\begin{pmatrix}u_n(i)\\ v_n(i)\end{pmatrix} \quad (6)$$

where $\hat{\xi}$ is the single-particle part of the hamiltonian (4). Equations (5) and (6) are solved self-consistently to determine Δ_i.

Including phase fluctuations. The Bogoliubov–de Gennes approximation completely neglects phase fluctuations of the order parameter, due to its mean-field nature. To account for thermal phase fluctuations, we ignore quantum fluctuations, that is, the time dependence of Δ in the partition function (3). The resulting partition function is:

$$Z=\int\prod_i\mathrm{d}|\Delta_i|\,\mathrm{d}\theta_i\exp\left(-\frac{\beta}{2U}\sum_i|\Delta_i|^2\right)\mathrm{Tr}\exp(-\beta H_{\mathrm{BdG}}) \quad (7)$$

where H_{BdG} is the Bogoliubov–de Gennes (BdG) hamiltonian (4), and so the partition function reads:

$$Z=\int\prod_i\mathrm{d}|\Delta_i|\,\mathrm{d}\theta_i\exp\left(-\frac{\beta}{2U}\sum_i|\Delta_i|^2\right)\prod_{n=1}^{2N}(1+\exp(-\beta E_n)) \quad (8)$$

where E_n are the eigenvalues of H_{BdG}.

The evaluation of expectation values and correlation functions for this partition function is carried out numerically using a Monte Carlo scheme[19,20]: at each step, a set of values $\{|\Delta_i|,\theta_i\}_{i=1}^{N}$ is chosen, inserted into H_{BdG}, which is then diagonalized. The integrand of equation (8) is then evaluated and weighted with temperature. However, for low enough temperatures such that the Monte Carlo averages of $|\Delta_i|$ hardly differ from those obtained from their mean-field values, one may take $|\Delta_i|$ in equation (7) to be their mean-field values, and the integral runs over the phases only. The phase correlations $\langle\cos(\delta\theta_i-\delta\theta_j)\rangle$ are then evaluated by:

$$\langle\cos(\delta\theta_i-\delta\theta_j)\rangle=\frac{1}{Z}\int\prod_i\mathrm{d}|\Delta_i|\,\mathrm{d}\theta_i\cos(\delta\theta_i-\delta\theta_j)\exp\left(-\frac{\beta}{2U}\sum_i|\Delta_i|^2\right)\prod_{n=1}^{2N}(1+\exp(-\beta E_n)) \quad (9)$$

where at each Monte Carlo step only the phases θ_i are changed and each phase configuration is given its thermal weight according to equation (9).

Received 16 May; accepted 15 August 2007.

1. Anderson, P. W. Theory of dirty superconductors. *J. Phys. Chem. Solids* **11,** 26–30 (1959).
2. Goldman, A. M. & Markovic, N. Superconductor-insulator transitions in the two-dimensional limit. *Phys. Today* **51,** 39–44 (1998).
3. Reich, S. *et al.* Localized high-T_c superconductivity on the surface of Na-doped WO 3. *J. Superconductivity* **13,** 855–861 (2000).
4. Cren, T., Roditchev, D., Sacks, W. & Klein, J. Nanometer scale mapping of the density of states in an inhomogeneous superconductor. *Europhys. Lett.* **54,** 84–90 (2001).
5. Pan, S. H. *et al.* Microscopic electronic inhomogeneity in the high-T_c superconductor $Bi_2Sr_2CaCu_2O_{8+x}$. *Nature* **413,** 282–285 (2001).
6. Kowal, D. & Ovadyahu, Z. Disorder induced granularity in an amorphous superconductor. *Solid State Commun.* **90,** 783–786 (1994).
7. Crane, R. W. *et al.* Survival of superconducting correlations across the two-dimensional superconductor-insulator transition: A finite-frequency study. *Phys. Rev. B* **75,** 184530 (2007).
8. Paalanen, M. A., Hebard, A. F. & Ruel, R. R. Low-temperature insulating phases of uniformly disordered two-dimensional superconductors. *Phys. Rev. Lett.* **69,** 1604–1607 (1992).
9. Gantmakher, V. F., Golubkov, M. V., Lok, J. G. S. & Geim, A. K. Giant negative magnetoresistance of semi-insulating amorphous indium oxide films in strong magnetic fields. *J. Exp. Theor. Phys.* **82,** 951–958 (1996).
10. Sambandamurthy, G., Engel, L. W., Johansson, A. & Shahar, D. Superconductivity-related insulating behavior. *Phys. Rev. Lett.* **92,** 107005 (2004).
11. Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. *Rep. Prog. Phys.* **62,** 61–122 (1999).
12. Alvarez, G., Mayr, M., Moreo, A. & Dagotto, E. Areas of superconductivity and giant proximity effects in underdoped cuprates. *Phys. Rev. B* **71,** 014514 (2005).
13. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. Theory of superconductivity. *Phys. Rev.* **108,** 1175–1204 (1957).
14. Ma, M. & Lee, P. A. Localized superconductors. *Phys. Rev. B* **32,** 5658–5667 (1985).
15. Fisher, M. P. A., Grinstein, G. & Girvin, S. M. Presence of quantum diffusion in two dimensions: Universal resistance at the superconductor-insulator transition. *Phys. Rev. Lett.* **64,** 587–590 (1990).
16. Fisher, M. P. A. Quantum phase transitions in disordered two-dimensional superconductors. *Phys. Rev. Lett.* **65,** 923–926 (1990).
17. Galitski, V. M. & Larkin, A. I. Disorder and quantum fluctuations in superconducting films in strong magnetic fields. *Phys. Rev. Lett.* **87,** 087001 (2001).
18. Ghosal, A., Randeria, M. & Trivedi, N. Role of spatial amplitude fluctuations in highly disordered s-wave superconductors. *Phys. Rev. Lett.* **81,** 3940–3943 (1998).
19. De-Gennes P. G. *Superconductivity of Metals and Alloys* (W. A. Benjamin, New York, 1966).
20. Mayr, M., Alvarez, G., Sen, C. & Dagotto, E. Phase fluctuations in strongly coupled d-wave superconductors. *Phys. Rev. Lett.* **94,** 217001 (2005).
21. Dubi, Y., Meir, Y. & Avishai, Y. Theory of magneto-resistance in disordered superconducting films. *Phys. Rev. B* **73,** 054509 (2006).
22. Clogston, A. M. Upper limit for the critical field in hard superconductors. *Phys. Rev. Lett.* **9,** 266–267 (1962).
23. Chandrasekhar, B. S. A note on the maximum critical field of high-field superconductors. *Appl. Phys. Lett.* **1,** 7–8 (1962).
24. Ephron, D., Yazdani, A., Kapitulnik, A. & Beasley, M. R. Observation of quantum dissipation in the vortex state of a highly disordered superconducting thin film. *Phys. Rev. Lett.* **76,** 1529–1532 (1996).
25. Aubin, H. *et al.* Magnetic-field-induced quantum superconductor-insulator transition in $Nb_{0.15}Si_{0.85}$. *Phys. Rev. B* **73,** 094521 (2006).
26. Hebard, A. F. & Paalanen, M. A. Magnetic-field-tuned superconductor-insulator transition in two-dimensional films. *Phys. Rev. Lett.* **65,** 927–930 (1990).

27. Yazdani, A. & Kapitulnik, A. Superconducting-insulating transition in two-dimensional a-MoGe thin films. *Phys. Rev. Lett.* **74,** 3037–3040 (1995).
28. Das Gupta, K., Sambandamurthy, G., Soman, S. S. & Chandrasekhar, N. Possible robust insulator-superconductor transition on solid inert gas and other substrates. *Phys. Rev. B* **63,** 104502 (2001).
29. Baturina, T. I. *et al.* Superconductivity on the localization threshold and magnetic-field-tuned superconductor-insulator transition in TiN films. *JETP Lett.* **79,** 337–341 (2004).
30. Emery, V. J. & Kivelson, S. A. Importance of phase fluctuations in superconductors with small superfluid density. *Nature* **374,** 434–437 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge discussions with A. Auerbach. This work was carried out with the support of the Israel Science Foundation and the US-Israel Binational Science Foundation. Y.D. acknowledges support from a Kreitman fellowship. Y.M. acknowledges the hospitality of the Aspen Center of Physics. Y.A. acknowledges JSPS fellowship.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.M. (ymeir@bgu.ac.il).

LETTERS

Nucleation and growth mechanism of ferroelectric domain-wall motion

Young-Han Shin[1]†, Ilya Grinberg[1], I-Wei Chen[2] & Andrew M. Rappe[1]

The motion of domain walls is critical to many applications involving ferroelectric materials, such as fast high-density non-volatile random access memory[1]. In memories of this sort, storing a data bit means increasing the size of one polar region at the expense of another, and hence the movement of a domain wall separating these regions. Experimental measurements of domain growth rates in the well-established ferroelectrics $PbTiO_3$ and $BaTiO_3$ have been performed, but the development of new materials has been hampered by a lack of microscopic understanding of how domain walls move[2–11]. Despite some success in interpreting domain-wall motion in terms of classical nucleation and growth models[12–16], these models were formulated without insight from first-principles-based calculations, and they portray a picture of a large, triangular nucleus that leads to unrealistically large depolarization and nucleation energies[5]. Here we use atomistic molecular dynamics and coarse-grained Monte Carlo simulations to analyse these processes, and demonstrate that the prevailing models are incorrect. Our multi-scale simulations reproduce experimental domain growth rates in $PbTiO_3$ and reveal small, square critical nuclei with a diffuse interface. A simple analytic model is also proposed, relating bulk polarization and gradient energies to wall nucleation and growth, and thus rationalizing all experimental rate measurements in $PbTiO_3$ and $BaTiO_3$.

Symmetry breaking by an external electric field on a 180° ferroelectric domain wall leads to domain-wall motion. The domain-wall speed has been found to be proportional to $\exp(-E_a/E)$ (Merz's law), where E_a is the 'activation field' and E is the applied electric field[2]. A classic theory supporting Merz's law was developed by Miller and Weinreich[15]. They suggested that the critical nucleus is an atomically thin triangular plate with a large aspect ratio, which then expands laterally on the same atomic plane[15]. However, the experimental and theoretical observations have shown that the Miller–Weinreich theory overestimates the activation field by an order of magnitude[5,17]. Thus, the details of the intrinsic properties and the mechanism of the domain dynamics are still unclear. This is due to experimental limitations in detecting rapid polarization changes in small regions and the computational difficulty of accurately simulating a sufficiently large supercell. Therefore, there is currently a gap between experimental measurements that probe domain-wall propagation (the result of nucleation and growth), and classical theories that focus on nucleation. This has motivated us to study domain-wall motion with molecular dynamics simulations[18,19] and for longer times and a larger length scale we model it with coarse-grained Monte Carlo simulations. This approach enables us to obtain the domain-wall speed in $PbTiO_3$ without the effect of defects and grain boundaries, giving the upper bound of the wall speed of real thin films.

Molecular dynamics simulations allow us precisely to isolate and study the nucleation aspect of the domain-wall motion. When an external field is applied (Fig. 1a and 1b), critical nuclei randomly form on the domain wall. A critical nucleus corresponds to the transition state between the reactant (the domain-wall layer polarized negatively, opposite to the applied field direction) and the product (the domain-wall layer polarized positively, along the applied field direction). This state of the system is found by analysing the trajectories of the positively polarized nuclei that appear on the domain wall; for a critical nucleus, the probabilities of nucleus growth and disappearance are the same. Nucleation was found to behave as a Poisson process, in that the probability that no critical nucleus has formed exponentially decreases with time t (Fig. 1c). This permits the extraction of a nucleation rate J from the molecular dynamics data.

For small supercells, interaction between periodic images may give rise to artefacts. To eliminate this problem, we carried out a series of domain-wall motion studies on supercells, increasing the wall area from $(3\times3)ac$ to $(6\times6)ac$, where a and c are the lattice constants of tetragonal $PbTiO_3$. The third dimension was kept constant at $18a$. For all temperatures, the nucleation rate converges when reaching $(6\times6)ac$ wall area (Fig. 1d). The rapid convergence of the nucleation rate J with cell size allows us to determine the simulation-area-independent, steady-state nucleation rate $J(T,E)$ for various temperatures T and applied electric fields E as presented in Table 1. Analysis of these results reveals that nucleation follows Merz's law with an activation field $E_{a,n}$ of 1.2–6.7 MV cm^{-1} over the 200–300 K temperature range.

Molecular dynamics simulations also show that the growth of critical nuclei is two-dimensional, with activation barriers much smaller than for nucleation. We identify six growth rates $G_{m,n}$, distinguished by the polarization of the four nearest neighbours around each five-atom unit cell on the domain wall. In this notation, m (or n) denotes the number of sideways (or forward) neighbours whose polarity is the same as the field direction (Fig. 2a). These growth rates are extracted from simulations with large domain-wall areas by digitizing the local polarizations on the domain wall, treating all cells with $P_z \geq 0$ as positively polarized neighbours and $P_z < 0$ as negatively polarized neighbours. We find that growth is also an activated process following Merz's law, with an activation field $E_{a,g}$ around 0.6–1.1 MV cm^{-1} over the 200–300 K temperature range for the slowest growth rate $G_{1,0}$ (Table 1).

Combining the molecular dynamics results for nucleation and growth processes, we used coarse-grained Monte Carlo simulations to model domain-wall propagation. We define the overall domain-wall speed v as the rate of increase in up-polarized domain volume divided by the wall area of the initial up-polarized domain. The obtained values for the total domain-wall speed v are shown in Fig. 2b. The speed also follows Merz's law with an activation field $E_{a,t}$ of 0.8–3.2 MV cm^{-1} over the 200–300 K temperature range. At lower field strengths, we find good agreement between our domain-wall velocities and room-temperature experimental data[5,8,11]. Additionally, the activation fields

[1]The Makineni Theoretical Laboratories, Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104–6323, USA. [2]Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104–6272, USA. †Present address: Department of Materials Science and Engineering, Pohang University of Science and Technology, Pohang 790–784, Korea.

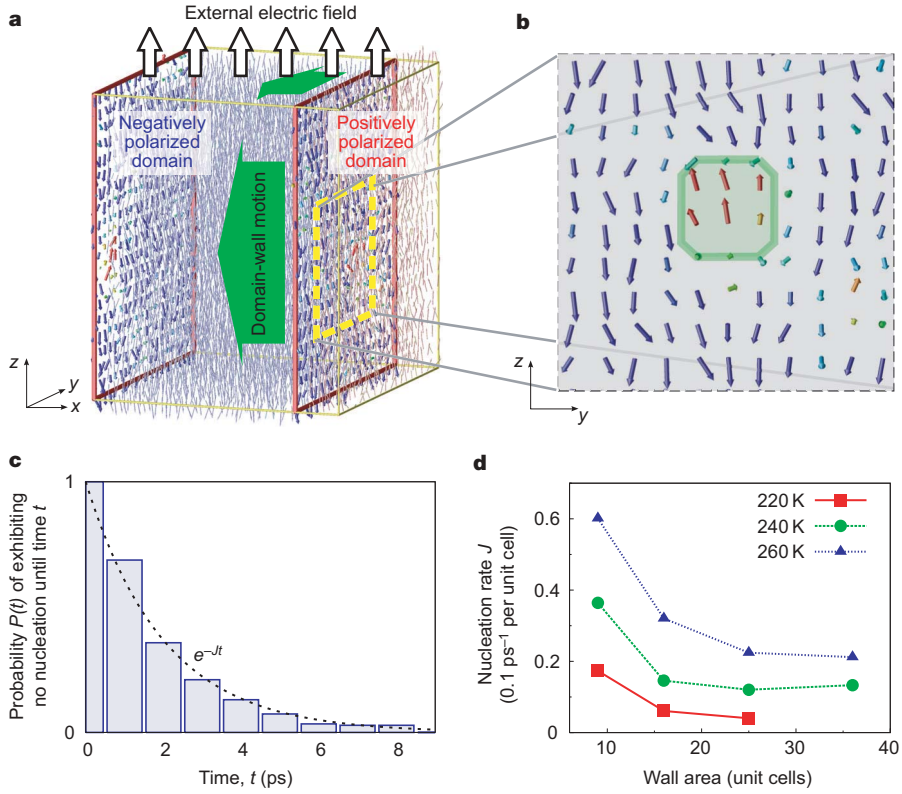


Figure 1 | Molecular dynamics simulations of nucleation on 180° domain walls. **a**, A snapshot of the polarization from a molecular dynamics simulation of $PbTiO_3$ at 220 K and 0.5 MV cm^{-1}. The local polarization vectors are shown by the red and blue arrows. The two domain walls are outlined with red lines. **b**, A critical nucleus on the domain wall on the *y*–*z* plane. The green solid line shows the boundary of the critical nucleus. **c**, The fraction of the simulations exhibiting no nucleation by time *t*. The exponent of the fit function corresponds to the nucleation rate *J*. **d**, Size dependence of the nucleation rates on the (100) 180° domain wall.

$E_{a,t}$ are in excellent accord with the experimental results of Tybell *et al.*[5] for $Pb(Zr_{0.2}Ti_{0.8})O_3$ films (~1.0 MV cm^{-1} at room temperature). The activation field of the overall wall velocity also agrees with the Avrami theory of transformation kinetics, which predicts the overall activation field to be roughly the weighted average of the nucleation and growth activation fields. That is, $E_{a,t} = 1/(d+1)E_{a,n} + d/(d+1)E_{a,g}$, where d is dimensionality, which is approximately 2 in our case[13].

For the nucleation process, our results are in gross disagreement with the Miller–Weinreich theory predictions (using $P_s = 0.76$ C m^{-2}, dielectric constant $\varepsilon = 60$ and a (100) domain-wall energy $\sigma_{100} = 0.11$ J m^{-2}): critical width $l_y = 43$ Å, critical height $l_z = 53$ Å and $E_{a,n} = 40$ MV cm^{-1} at $E = 0.5$ MV cm^{-1} and $T = 240$ K. Such a high activation field would lead to a domain-wall velocity many orders of magnitude smaller than that observed in experiments or our simulations. This reconfirms the well-known disagreement between the Miller–Weinreich predictions of the activation energy and field and the experimental observations. Detailed examination of the critical nuclei using large 20 × 20 supercells reveals further significant discrepancies between our molecular dynamics results and the traditional Miller–Weinreich model of nucleation. A snapshot of the polarization reversal process (Fig. 1b) portrays a 12 Å × 12 Å critical nucleus that is nearly square, not triangular in shape. Figure 1b also shows considerable diffuseness, in contrast to the sharp polarization reversal in the nucleus assumed by the Miller–Weinreich theory.

We therefore develop an analytic model to relate computed and experimental domain-wall velocities to material properties that are easily obtained from bulk experimental data or from static density functional theory (DFT) calculations. We base our nucleation model on the Landau–Ginzburg–Devonshire approach. Here, the structure at the domain wall is due to interplay between two energy terms. The first term reflects the local energy cost for the polarization to deviate from the spontaneous polarization P_s

$$U_{loc}(P_z) = A_{loc}\left[1 - \left(\frac{P_z}{P_s}\right)^2\right]^2 \quad (1)$$

where the constant A_{loc} is 3.55×10^8 J m^{-3} at 0 K for our atomistic model and decreases with temperature as $P_s^4(T)$. This term is zero for a sharp polarization reversal from P_s to $-P_s$, but rises with diffuseness. The second term represents the preference of electric dipoles to align, associating higher gradient energy with larger polarization gradients

$$U_m(P_z) = g_m\left(\frac{\partial P_z}{\partial m}\right)^2 \quad (2)$$

Table 1 | Nucleation and growth rates and activation fields

T (K)	*E* = 0.45 MV cm^{-1}		*E* = 0.50 MV cm^{-1}		*E* = 0.55 MV cm^{-1}		*E* = 0.60 MV cm^{-1}		*E* = 0.65 MV cm^{-1}		Activation field		
	J	$G_{1,0}$	*J*	$G_{1,0}$	*J*	$G_{1,0}$	*J*	$G_{1,0}$	*J*	$G_{1,0}$	$E_{a,n}$	$E_{a,g}$	$E_{a,t}$
200					1.9	12.4	5.8	13.7			6.7	1.1	3.2
220			3.2	13.2	8.5	14.7	18.2	16.4	29.9	20.1	4.9	0.9	2.1
240	5.0	12.3	11.0	15.8	22.5	17.6	33.6	20.9	47.6	23.3	3.3	0.9	1.5
260	14.7	16.9	25.2	18.4	38.8	20.6	50.0	21.4	62.5	24.0	1.9	0.5	1.0
280	27.6	17.8	40.4	19.0	51.3	22.9	63.8	23.9			1.5	0.6	0.9
300	40.2	21.0	55.2	23.7	65.9	25.7	77.7	29.0			1.2	0.6	0.8

The nucleation rate *J* is measured in 10^{-3} ps^{-1} per unit cell and the growth rate $G_{1,0}$ is measured in ps^{-1} per unit cell.

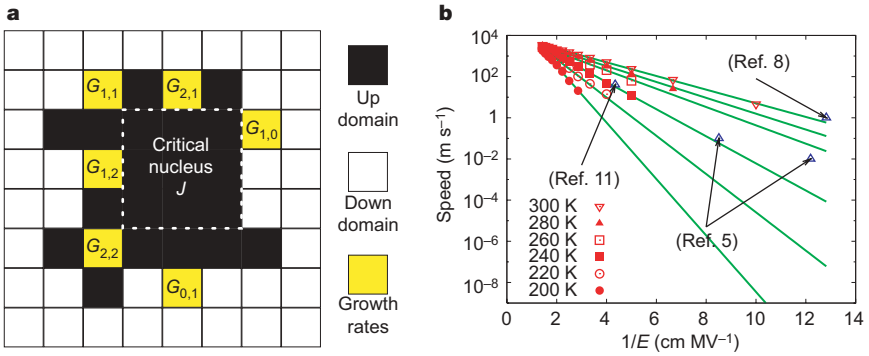


Figure 2 | Coarse-grained Monte Carlo simulations of polarization switching. a, Definition of growth rates. $G_{m,n}$ means the rate of the event of changing the polarization of a unit cell, which has m sideways neighbours and n forward neighbours polarized along the field direction. Here the direction of the field is 'up'. Growth rates $G_{m,n}$ can be deduced from counting the required time for flipping yellow unit cells. **b**, The overall domain-wall speeds as a function of T and E. They are comparable to recent experiments and conform to Merz's law: speed $v \approx \exp(-E_{a,t}/E)$.

where g_m is the gradient coefficient along the m direction. According to our molecular dynamics simulations, the coefficient along the polar axis g_z is 1.07×10^{11} m^3 F^{-1}, and the coefficients normal to the polar axis g_x and g_y are 0.63×10^{-11} m^3 F^{-1} (see Methods). The energy difference ΔU in a two-dimensional nucleus on the domain wall relative to the domain wall without any nucleus can be expressed as a function of its size l_y and l_z, and the z-component of the local polarization $P_z(x, y, z)$:

$$\Delta U = \Delta U_v + \Delta U_i \quad (3)$$

$$\Delta U_v = -E \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dz \left(P_z(x, y, z) - P_z^{180}(x, y, z)\right) \quad (4)$$

$$\Delta U_i = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dz \left\{ \left[U_x(P_z) + U_y(P_z) + U_z(P_z) + U_{loc}(P_z)\right] - \left[U_x(P_z^{180}) + U_{loc}(P_z^{180})\right] \right\} \quad (5)$$

Here P_z^{180} is the polarization of the clean 180° domain wall without any nucleus on it, and the reversal of $P_z(x, y, z)$ around the nucleus is spread across a diffuse interface with the lateral diffuseness parameters δ_y and δ_z and the transverse diffuseness parameter δ_x. The critical nucleus is then obtained by numerically locating the saddle point of ΔU. Using the polarization and A_{loc} values appropriate for our model potential, we find $l_y = 12$ Å, $l_z = 12$ Å, $\delta_y = 3.9$ Å, $\delta_z = 4.6$ Å at $T = 240$ K and $E = 0.5$ MV cm^{-1} (Supplementary Fig. 5a), in agreement with the microscopic nucleus structure shown in Fig. 1b. We also obtain the activation fields $E_{a,n}$ in excellent agreement with simulation results for the whole 200–300 K range of simulations (Fig. 3c). We further obtain $\delta_x = 3.5$ Å, indicating the nucleus is still very much limited to one atomic plane despite lateral diffuseness.

The much smaller l_y, l_z and $E_{a,n}$ are due to three new features, which are absent in the Miller–Weinreich model. First, the amount of the interface area for any nucleus is significantly less than the Miller–Weinreich model estimate due to lateral diffuseness. In the Miller–Weinreich model (Fig. 3a), nucleation creates additional domain-wall area of Sa (lattice constants $a = 3.9$ Å and $c = 4.15$ Å, $c/a \approx 1$, and S is the nucleus perimeter). The domain wall passes through all atoms with $P_z = 0$, and between neighbouring up- and down-polarized atoms. In the new model, Pb off-centring (A–D in Fig. 3b) allows a slanted (101) domain wall of area $Sa\sqrt{2}$ to replace the flat wall area (Sa), resulting in net wall creation of $Sa(\sqrt{2} - 1) \approx 0.4Sa$. Second, in contrast to full up- and down-polarizations at two adjacent sites in the Miller–Weinreich model, our model places smaller polarizations at these sites, reducing the polarization gradient, and hence the gradient energy, by 50%. The energy per unit area of the interface is lower than in the Miller–Weinreich model because polarization changes with the finite diffuseness parameters δ_y and δ_z, and P_z is smaller than P_s. Taken together, these effects show that for the same nucleus perimeter, the interface energy cost is a factor of three to four lower than that estimated by the Miller–Weinreich theory, directly leading to an order-of-magnitude reduction in critical nucleus area and energy.

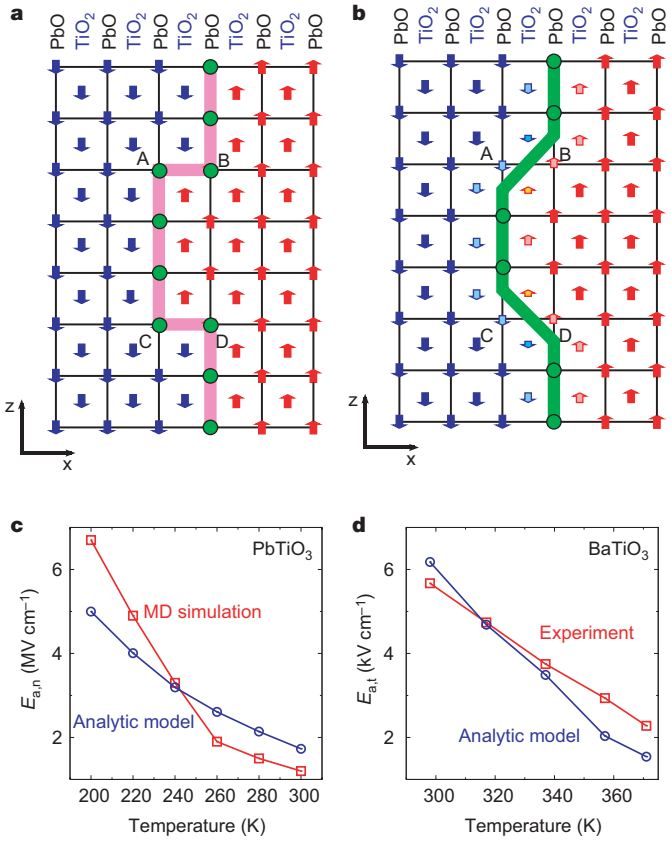


Figure 3 | Landau-Ginzburg-Devonshire model of nucleation on 180° domain walls. a, **b**, Schematic diagrams of the Miller–Weinreich model (**a**) and the diffuse-boundary model (**b**) on the x–z plane, with $P_z = 0$ marked by circles and Pb off-centring at A–D. In the Miller–Weinreich picture, all the Pb atoms at the nucleus boundary (circled in **a**) have no off-centre displacement, as shown by the lack of arrows along the z direction. In the Landau–Ginzburg–Devonshire model, the edge of the nucleus boundary bypasses the Pb atoms (A–D in **b**) that are polarized in the asymmetric environment. **c**, Comparison of the activation fields obtained from simulations of $PbTiO_3$ (squares) with the results of the Landau–Ginzburg–Devonshire model (circles). **d**, Comparison of the activation fields obtained experimentally (ref. 22) for $BaTiO_3$ (squares) with the results of the Landau–Ginzburg–Devonshire model (circles).

Third, since the critical nucleus is small, the depolarization energy is not important during the nucleation process. This enables the formation of square nuclei instead of triangular ones.

Because our model can predict $E_{a,n}$ without any molecular dynamics simulations, we have extended it to $BaTiO_3$ using only experimental P_s and energies obtained by DFT calculations. In our theory, four parameters are necessary for modelling nucleation at a given temperature T: g_x, g_z, A_{loc} and $P_s(T)$. From five-atom bulk $BaTiO_3$ DFT calculations with the local density approximation in the literature, we obtain $A_{loc} = 9.5$ meV per unit cell[20]. The 180° (100) domain-wall energy is $\sigma_{100} = 6$–$7.5\ mJ\ m^{-2}$ according to the DFT calculations in refs 17 and 21. The P_s of tetragonal $BaTiO_3$ is $0.31\ C\ m^{-2}$ at 0 K and its temperature dependence is given by ref. 22. From these data, we can obtain $g_x = 0.22 \times 10^{-11}\ m^3\ F^{-1}$ (see Methods). Using this g_x value for $BaTiO_3$ and the ratio between g_x and g_z determined for $PbTiO_3$, we estimate $g_z = 0.38 \times 10^{-11}\ m^3\ F^{-1}$. Our Landau–Ginzburg–Devonshire model then predicts $E_{a,n}$ values of 4.6–18.5 kV cm^{-1} for the 298–371 K temperature range. According to the Avrami equation, $E_{a,t}$ should be larger than $E_{a,n}/3$. This allows us to obtain a lower bound for $E_{a,t}$ values that compares well with the experimental data of ref. 22, as shown in Fig. 3d, indicating very fast growth in $BaTiO_3$ in these experiments. In contrast, using the DFT-obtained domain-wall energy, the Miller–Weinreich theory would predict $E_{a,n} = 240\ kV\ cm^{-1}$ for $T = 300$ K, which is clearly out of the question.

In conclusion, our multi-scale modelling finds domain-wall velocities in agreement with the most direct experimental evidence for related materials. It also allows the construction of a general and accurate model for a critical nucleus. Together, they provide microscopic insight into how the changes in the nucleus shape and polarization profile can dramatically lower activation barriers.

METHODS SUMMARY

We studied many $PbTiO_3$ structures with first-principles DFT, and these provided a database of energies and forces for calibrating an inter-atomic potential which was based on Brown's rules of valence[23]. The molecular dynamics simulations exhibited fundamental domain-wall processes, including critical nucleus formation and growth. Initially, three layers were polarized up and 15 layers were polarized down, and each layer had $N_y \times N_z$ unit cells. After equilibrating the system, we applied the electric field in the range of 0.45 to 0.65 MV cm^{-1} under the Nosé–Hoover thermostat with a 1 fs time step. Nucleation is a stochastic process, so nucleation rates were obtained from the statistics of molecular dynamics simulations. Two hundred randomly dispersed initial coordinates were used to determine each nucleation rate. To model larger-scale, longer-time domain-wall motion processes, we simulated phase transformation of ferroelectric domains with the nucleation-and-growth model[13,19,24–26]. The local polarization of each primitive five-atom $PbTiO_3$ unit cell was chosen as a unit in this coarse-grained Monte Carlo simulation, while the nucleation and growth rates were obtained from the analysis of the molecular dynamics simulations.

The overall domain-wall speed v can be obtained from the Monte Carlo study using the following relation $v = \frac{1}{NA}\frac{\partial V_{up}}{\partial t} = \frac{d_1}{N}\frac{\partial}{\partial t}\sum_l q_l(t)$, where $V_{up} = Ad_1\sum_l q_l(t)$ is the up-polarized domain volume, A is the wall area, $d_1 \approx a$ is the distance between adjacent layers, q_l is the normalized polarization fraction of the lth layer, and N is the number of domain walls ($N = 2$ owing to the periodic boundary conditions in our coarse-grained Monte Carlo simulations). Such a multiscale strategy enabled us to increase the system size to the micrometre scale and accurately evaluate the overall domain-wall-motion speed.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 April; accepted 6 August 2007.
Published online 7 October 2007.

1. Scott, J. F. & Paz de Araujo, C. A. Ferroelectric memories. *Science* **246,** 1400–1405 (1989).
2. Merz, W. J. Domain formation and domain wall motions in ferroelectric $BaTiO_3$ single crystals. *Phys. Rev.* **95,** 690–698 (1954).
3. Stadler, H. L. & Zachmanidis, P. J. Nucleation and growth of ferroelectric domains in $BaTiO_3$ at fields from 2 to 450 kV/cm. *J. Appl. Phys.* **34,** 3255–3260 (1963).
4. Ganpule, C. S. *et al.* Role of 90° domains in lead zirconate titanate thin films. *Appl. Phys. Lett.* **77,** 292–294 (2000).
5. Tybell, T., Paruch, P., Giamarchi, T. & Triscone, J.-M. Domain wall creep in epitaxial ferroelectric $Pb(Zr_{0.2}Ti_{0.8})O_3$ thin films. *Phys. Rev. Lett.* **89,** 097601 (2002).
6. Ahn, C. H., Rabe, K. M. & Triscone, J.-M. Ferroelectricity at the nanoscale: local polarization in oxide thin films and heterostructures. *Science* **303,** 488–491 (2004).
7. Li, J. *et al.* Ultrafast polarization switching in thin-film ferroelectrics. *Appl. Phys. Lett.* **84,** 1174–1176 (2004).
8. Gruverman, A. *et al.* Direct studies of domain switching dynamics in thin film ferroelectric capacitors. *Appl. Phys. Lett.* **87,** 082902 (2005).
9. So, Y. W., Kim, D. J., Noh, T. W., Yoon, J.-G. & Song, T. K. Polarization switching kinetics of epitaxial $Pb(Zr_{0.4}Ti_{0.6})O_3$ thin films. *Appl. Phys. Lett.* **86,** 092905 (2005).
10. Stolichnov, I., Malin, L., Colla, E., Tagantsev, A. K. & Setter, N. Microscopic aspects of the region-by-region polarization reversal kinetics of polycrystalline ferroelectric $Pb(Zr,Ti)O_3$ films. *Appl. Phys. Lett.* **86,** 012902 (2005).
11. Grigoriev, A. *et al.* Nanosecond domain wall dynamics in ferroelectric $Pb(Zr,Ti)O_3$ thin films. *Phys. Rev. Lett.* **96,** 187601 (2006).
12. Landauer, R. Electrostatic considerations in $BaTiO_3$ domain formation during polarization reversal. *J. Appl. Phys.* **28,** 227–234 (1957).
13. Shur, V., Rumyantsev, E. & Makarov, S. Kinetics of phase transformations in real finite systems: application to switching in ferroelectrics. *J. Appl. Phys.* **84,** 445–451 (1998).
14. Hayashi, M. Kinetics of domain wall motion in ferroelectric switching. I. General formation. *J. Phys. Soc. Jpn* **33,** 616–628 (1972).
15. Miller, R. C. & Weinreich, G. Mechanism for the sidewise motion of 180° domain walls in barium titanate. *Phys. Rev.* **117,** 1460–1466 (1960).
16. Orihara, H., Hashimoto, S. & Ishibashi, Y. A theory of *D-E* hysterisis loop based on the Avrami model. *J. Phys. Soc. Jpn* **63,** 1031–1035 (1994).
17. Padilla, J., Zhong, W. & Vanderbilt, D. First-principles investigation of 180° domain walls in $BaTiO_3$. *Phys. Rev. B* **53,** R5969–R5973 (1996).
18. Grinberg, I., Cooper, V. R. & Rappe, A. M. Relationship between local structure and phase transitions of a disordered solid solution. *Nature* **419,** 909–911 (2002).
19. Shin, Y.-H., Cooper, V. R., Grinberg, I. & Rappe, A. M. Development of a bond-valence molecular-dynamics model for complex oxides. *Phys. Rev. B* **71,** 054104 (2005).
20. Cohen, R. E. Origin of ferroelectricity in perovskite oxides. *Nature* **358,** 136–138 (1992).
21. Meyer, B. & Vanderbilt, D. *Ab initio* study of ferroelectric domain walls in $PbTiO_3$. *Phys. Rev. B* **65,** 104111 (2002).
22. Savage, A. & Miller, R. C. Temperature dependence of the velocity of sidewise 180° domain-wall motion in $BaTiO_3$. *J. Appl. Phys.* **31,** 1546–1549 (1960).
23. Brown, I. D. & Wu, K. K. Empirical parameters for calculating cation-oxygen bond valences. *Acta Crystallogr.* **B32,** 1957–1959 (1976).
24. Avrami, M. Kinetics of phase change. I. General theory. *J. Phys. Chem.* **7,** 1103–1112 (1939).
25. Kashchiev, D. *Nucleation: Basic Theory with Applications* Ch. 26 (Butterworth-Heinemann, Woburn, Massachusetts, 2000).
26. Lines, M. E. & Glass, A. M. *Principles and Applications of Ferroelectrics and Related Materials* Ch. 4 (Clarendon Press, Oxford, 1977).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This material is based upon work supported by the US Office of Naval Research, the National Science Foundation and the Army Engineer Research and Development Center. Computational support was provided by the US Department of Defense. Y.-H.S. was supported by the Brain Korea 21 project in 2006.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.M.R. (rappe@sas.upenn.edu).

METHODS

Here we explain the details of the analytical model of the critical nucleus on the ferroelectric domain wall, as well as our procedure for extracting the model coefficients from molecular dynamics simulations. The model discussed here closely follows the treatment in ref. 26.

Total free energy. The total free energy difference of a stripe domain is:

$$G - G_0 = W_g + W_E + W_e + W_l \tag{6}$$

where G_0 is the single domain free energy, W_g is the gradient energy due to the dipole-dipole interaction, W_E is the depolarizing energy, W_e is the elastic energy, and W_l is the local energy contribution from the Landau–Ginzburg–Devonshire phenomenological theory. Two minima of $E(P) = 1/2AP^2 + 1/4BP^4$ where $A = A_0(T - T_c)$, $A_0 > 0$, and $B > 0$ are located at $P_s = \pm\sqrt{(-A/B)}$, and the local energy W_l can be expressed as:

$$W_l = \iiint U_{loc}(P(x,y,z))\,dx dy dz \tag{7}$$

where $P(x, y, z)$ is the magnitude of a general polarization vector $\boldsymbol{P}(x, y, z)$, and the local energy per unit cell $U_{loc}(P)$ is:

$$\begin{aligned} U_{loc}(P(x,\,y,\,z;\;T)) &= E(P(x,\,y,\,z;\;T)) - E(P_s(T)) \\ &= \left(\frac{1}{2}AP^2 + \frac{1}{4}BP^4\right) - \left(\frac{1}{2}AP_s^2 + \frac{1}{4}BP_s^4\right) \\ &= \left(\frac{1}{2}(-BP_s^2)P^2 + \frac{1}{4}BP^4\right) - \left(\frac{1}{2}(-BP_s^2)P_s^2 + \frac{1}{4}BP_s^4\right) \\ &= A_{loc}(T)\left(1 - \left(\frac{P(x,\,y,\,z;\;T)}{P_s(T)}\right)^2\right)^2 \end{aligned} \tag{8}$$

where $A_{loc}(T) = -E(P_s(T)) = 1/4BP_s^4(T)$ is the energy difference per unit cell between the relaxed tetragonal ferroelectric phase and the high-symmetry tetragonal paraelectric phase, and P_s is the spontaneous polarization. For the 180° (100) domain wall, only the gradient energy W_g and the local energy W_l remain if we neglect the effect due to the free charge on the crystal surfaces.

Extracting model parameter from the (100) domain-wall energy. The 180° domain-wall energy σ_{100} is given by

$$\sigma_{100} = \frac{E_{100}^{2m\times1\times1} - E_{mono}^{2m\times1\times1}}{2A_{100}} = \frac{W_g + W_l}{2A_{100}} \tag{9}$$

where A_{100} is the area of the domain wall whose normal vector is [1 0 0] ($A_{100} = ac$ for the $2m \times 1 \times 1$ supercell, and a and c are the lattice constants ($a < c$)). $E_{100}^{2m\times1\times1}$ is the total energy of up–down domains, and $E_{mono}^{2m\times1\times1}$ is the total energy of a single domain. In the energy calculation, one supercell is composed of $2m \times 1 \times 1$ unit cells. Because we used periodic boundary conditions in this calculation, there are two domain walls in a supercell, which is the reason for the 2 in the denominator of equation (9). The spontaneous polarization P_s of $PbTiO_3$ is 0.89 C m^{-2} at 0 K, and the polarization next to the domain wall P_b is smaller than P_s; the diffuseness parameter (or domain-wall width) δ_x is approximately one unit cell. The polarization $P_z(x)$ across the domain wall looks like $-P_s, \ldots, -P_s, -P_b, 0, P_b, P_s, \ldots, P_s$, where $P_s = 0.89$ C m^{-2} and $P_b = 0.73$ C m^{-2} (Supplementary Table 2).

From the Landau–Ginzburg–Devonshire model with gradient terms, the polarization around the clean 180° (100) domain-wall boundary P_0 can be expressed with the hyperbolic function:

$$P_o(x,\,y,\,z) = P_s \tanh\left(\frac{x}{\delta_x/2}\right) = p(x) \tag{10}$$

which is shown in Supplementary Fig. 1. The parameter δ_x is the domain-wall width found by fitting the polarization data across the domain wall to equation (10).

Following equations (8) and (10), σ_{100} from the $2m \times 1 \times 1$ super cell can be expressed as:

$$\sigma_{100} = \left(W_l^{100} + W_{gx}^{100}\right)/(2A_{100}) \tag{11}$$

$$\begin{aligned} W_l^{100} &= \int_{-ma}^{ma} dx \int_{-a/2}^{a/2} dy \int_{-c/2}^{c/2} dz\; U_{loc}(P_{0,z}(x,\,y,\,z)) = ac \int_{-ma}^{ma} U_{loc}(p)dx \\ &= ac \int_{-P_s}^{P_s} \frac{dp}{\frac{2}{\delta_x}\frac{P_s^2 - p^2}{P_s}} A_{loc}\left(1 - \left(\frac{p}{P_s}\right)^2\right)^2 = \frac{acA_{loc}\delta_x}{2P_s^3}\int_{-P_s}^{P_s}\left(P_s^2 - p^2\right)dp \\ &= \frac{2}{3}acA_{loc}\delta_x \end{aligned} \tag{12}$$

$$\begin{aligned} W_{gx}^{100} &= g_x \int_{-c/2}^{c/2} dz \int_{-a/2}^{a/2} dy \int_{-ma}^{ma} dx \left(\frac{dp}{dx}\right)^2 \\ &= g_x \int_{-P_s}^{P_s} \left(\frac{dp}{dx}\right) dp = \frac{2g_x ac}{\delta_x P_s}\int_{-P_s}^{P_s}(P_s^2 - p^2)dp \\ &= \frac{8ac}{3\delta_x}P_s^2 g_x \end{aligned} \tag{13}$$

where the factor 2 in the denominator of equation (11) is due to two domain walls per supercell. By $\partial\sigma_{100}/\partial\delta_x\big|_{\delta_x = \delta_x^*} = 0$, the diffuseness parameter δ_x^* and the gradient coefficient g_x are

$$\delta_x^* = 2P_s\sqrt{\frac{g_x}{A_{loc}}} \tag{14}$$

$$\sigma_{100} = \frac{8P_s}{3}\sqrt{g_x A_{loc}} \tag{15}$$

$$g_x = \left(\frac{3\sigma_{100}}{8P_s}\right)^2 \frac{1}{A_{loc}} \tag{16}$$

σ_{100} and P_s are obtained from molecular dynamics simulations, and they are 0.125 J m^{-2} and 0.89 C m^{-2}, respectively. We use $A_{loc} = 3.55 \times 10^8$ J m^{-3} or 0.14 eV per unit cell for $PbTiO_3$ at 0 K. From these values and from equation (16), g_x equals 0.63×10^{-11} m^3 F^{-1}. By integrating equation (13) numerically with the polarization from molecular dynamics simulations, we obtain $g_x = 0.58 \times 10^{-11}$ m^3 F^{-1}. We attribute the small discrepancy between the two estimates to the difference between the continuous and discrete polarization.

Extracting model parameter from the (n01) domain-wall energy σ_{n01}. The 180° (n01) domain-wall energy σ_{n01} of the higher-index interface is:

$$\sigma_{n01} = \frac{E_{n01}^{2m\times1\times n} - E_{mono}^{2m\times1\times n}}{2A_{n01}} \tag{17}$$

where A_{n01} is the area of the $2m \times 1 \times n$ stepped domain wall whose normal vector is [n 0 1] and $A_{n01} = a\sqrt{a^2 + (nc)^2}$. The 180° ($n$01) domain wall is shown in Supplementary Fig. 2a. With an assumption that $P_x(x, y, z) = P_y(x, y, z) = 0$, σ_{n01} can be expressed as:

$$\sigma_{n01} = (W_l + W_{gx} + W_{gz})/(2A_{n01}) \tag{18}$$

$$W_l = \int_{-ma}^{ma} dx \int_{-a/2}^{a/2} dy \int_{-nc}^{nc} dz\; U_{loc}(P_z(x,\,y,\,z)) \tag{19}$$

$$W_{gx} = g_x \int_{-ma}^{ma} dx \int_{-a/2}^{a/2} dy \int_{-nc}^{nc} dz \left(\frac{\partial P_z(x,y,z)}{\partial x}\right)^2, \tag{20}$$

$$W_{gz} = g_z \int_{-ma}^{ma} dx \int_{-a/2}^{a/2} dy \int_{-nc}^{nc} dz \left(\frac{\partial P_z(x,\,y,\,z)}{\partial z}\right)^2 \tag{21}$$

where $\sigma_{n01} = 0.117$ J m^{-2}. From the numerical calculations, we can determine the gradient coefficient g_z as a function of n. We find that g_z converges to 1.07×10^{-11} m^3 F^{-1} as n and m increase (Supplementary Fig. 2b).

Model of the nucleus on the domain wall. To describe the diffuse polarization profile around the nucleus on the domain wall, equation (10) can be generalized as:

$$\begin{aligned} P_z(x,\,y,\,z) &= P_s\left(f^-(x,\,l_x,\,\delta_x)\left(2\frac{f^-(y,\,l_y,\,\delta_y)f^-(z,\,l_z,\,\delta_z)}{f^-(0,\,l_y,\,\delta_y)f^-(0,\,l_z,\,\delta_z)} - 1\right) + f^+(x,\,l_x,\,\delta_x)\right) \\ &\approx 2P_s f^-(x,\,l_x,\,\delta_x)f^-(y,\,l_y,\,\delta_y)f^-(z,\,l_z,\,\delta_z) + P_z^{180}(x - l_x/2,\,y,\,z) \end{aligned} \tag{22}$$

where $f^{\pm}(\alpha, \beta, \gamma) = \frac{1}{2}\tanh\left(\frac{\alpha+\beta/2}{\gamma/2}\right) \pm \frac{1}{2}\tanh\left(\frac{\alpha-\beta/2}{\gamma/2}\right)$, l_k corresponds to the length of the nucleus to the k direction, and δ_k corresponds to the diffuseness parameter along the k direction. The polarization profile generated by equation (22) is shown in Supplementary Fig. 4. When the external field E is applied to the 180° domain wall, the free energy change ΔU from the formation of a nucleus is

$$\Delta U = \Delta U_v + \Delta U_i \tag{23}$$

$$\Delta U_{\mathrm{v}} = -E \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}y\mathrm{d}z \left(P_z(x, y, z) - P_z^{180}(x, y, z)\right) \tag{24}$$

$$\Delta U_{\mathrm{i}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}x\mathrm{d}y\mathrm{d}z \left\{ \left[g_x\left(\frac{\partial P_z}{\partial x}\right)^2 + g_y\left(\frac{\partial P_z}{\partial y}\right)^2 + g_z\left(\frac{\partial P_z}{\partial z}\right)^2 + U_{\mathrm{loc}}(P_z)\right] - \left[g_x\left(\frac{\partial P_z^{180}}{\partial x}\right)^2 + U_{\mathrm{loc}}\left(P_z^{180}\right)\right]\right\} \tag{25}$$

where $U_{\mathrm{loc}}(p) = A_{\mathrm{loc}}\left(1-(p/p_{\mathrm{s}})^2\right)^2$ and the subscripts v and i mean volume and interface, respectively. We ignore the charge on the nucleus, which is justified, a posteriori, because the depolarization energy in the Miller–Weinreich model is strongly dependent on size, and is negligible when the nucleus is very small. Because we can also find the bulk polarization P_{s} from our model potential, the temperature dependence of $A_{\mathrm{loc}}(T)$ can be deduced from A_{loc} at 0 K and P_{s} at finite temperature.

Using the gradient parameters from the (100) and (n01) domain walls ($g_x = g_y = 0.63 \times 10^{-11}\,\mathrm{m^3\,F^{-1}}$ and $g_z = 1.07 \times 10^{-11}\,\mathrm{m^3\,F^{-1}}$), we can obtain the following results. The aspect ratio of the critical nucleus is close to 1. The critical diffuseness parameter along the polar axis δ_z^* is 4.6 Å, while the diffuseness parameter along the y axis δ_y^* is 3.5 Å. The size of the critical nucleus from this model is 3×3 unit cells and the activation barrier is 0.14 eV at 240 K and 500 kV cm^{-1}. This result is consistent with the molecular dynamics simulations, where we found that the activation energy is about 0.1 eV, and the critical nucleus was more diffuse along the polar axis direction than along the normal to the 180° domain wall.

LETTERS

Coaxial silicon nanowires as solar cells and nanoelectronic power sources

Bozhi Tian[1]*, Xiaolin Zheng[1]*, Thomas J. Kempa[1], Ying Fang[1], Nanfang Yu[2], Guihua Yu[1], Jinlin Huang[1] & Charles M. Lieber[1,2]

Solar cells are attractive candidates for clean and renewable power[1,2]; with miniaturization, they might also serve as integrated power sources for nanoelectronic systems. The use of nanostructures or nanostructured materials represents a general approach to reduce both cost and size and to improve efficiency in photovoltaics[1–9]. Nanoparticles, nanorods and nanowires have been used to improve charge collection efficiency in polymer-blend[4] and dye-sensitized solar cells[5,6], to demonstrate carrier multiplication[7], and to enable low-temperature processing of photovoltaic devices[3–6]. Moreover, recent theoretical studies have indicated that coaxial nanowire structures could improve carrier collection and overall efficiency with respect to single-crystal bulk semiconductors of the same materials[8,9]. However, solar cells based on hybrid nanoarchitectures suffer from relatively low efficiencies and poor stabilities[1]. In addition, previous studies have not yet addressed their use as photovoltaic power elements in nanoelectronics. Here we report the realization of p-type/intrinsic/n-type (p-i-n) coaxial silicon nanowire solar cells. Under one solar equivalent (1-sun) illumination, the p-i-n silicon nanowire elements yield a maximum power output of up to 200 pW per nanowire device and an apparent energy conversion efficiency of up to 3.4 per cent, with stable and improved efficiencies achievable at high-flux illuminations. Furthermore, we show that individual and interconnected silicon nanowire photovoltaic elements can serve as robust power sources to drive functional nanoelectronic sensors and logic gates. These coaxial silicon nanowire photovoltaic elements provide a new nanoscale test bed for studies of photo-induced energy/charge transport and artificial photosynthesis[10], and might find general usage as elements for powering ultralow-power electronics[11] and diverse nanosystems[12,13].

We have focused on p-i-n coaxial silicon nanowire structures (Fig. 1a) consisting of a p-type silicon nanowire core capped with i- and n-type silicon shells. An advantage of this core/shell architecture is that carrier separation takes place in the radial versus the longer axial direction, with a carrier collection distance smaller or comparable to the minority carrier diffusion length[8]. Hence, photogenerated carriers can reach the p-i-n junction with high efficiency without substantial bulk recombination. An additional consequence of this geometry is that material quality can be lower than in a traditional p-n junction device without causing large bulk recombination[1].

Silicon nanowire p-cores were synthesized by means of a nanocluster-catalysed vapour–liquid–solid (VLS) method[14,15]. Silicon shells were then deposited at a higher temperature and lower pressure than for p-core growth (Fig. 1a, right panel) to inhibit axial elongation of the silicon nanowire core during the shell deposition, where phosphine was used as the n-type dopant in the outer shell[15]. The growth temperatures were sufficiently low to ensure that minimal amounts of metal catalyst were incorporated into the silicon nanowire structure. Scanning electron microscopy (SEM) images of a typical p-i-n coaxial silicon nanowire recorded in the back-scattered electron mode (Fig. 1b) highlight several key features. First, the uniform contrast of the nanowire core is consistent with a single-crystalline structure expected for silicon nanowires obtained by the VLS method[14,15]. Second, contrast variation observed in the shells is indicative of a polycrystalline structure grain of the order of 30–80 nm. Third, the core/shell silicon nanowires have uniform

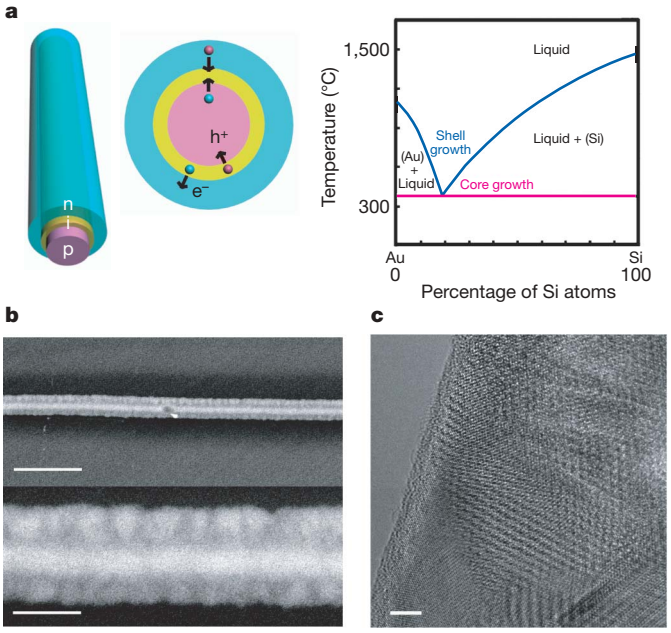


Figure 1 | Schematics and electron microscopy images of the p-i-n coaxial silicon nanowire. a, Illustrations of the core/shell silicon nanowire structure; its cross-sectional diagram shows that the photogenerated electrons (e^-) and holes (h^+) are swept into the n-shell and p-core, respectively, by the built-in electric field. The phase diagram of gold (Au)–silicon (Si) alloy on the right panel illustrates that the core is grown by means of the VLS mechanism, whereas the shells are deposited at higher temperature and lower pressure to inhibit further nanowire axial elongation. **b**, SEM images (back-scattered electron mode) of the p-i-n coaxial silicon nanowire at two different magnifications. Scale bar, 1 μm (top), 200 nm (bottom). The p-i-n silicon nanowire was grown with 100-nm-diameter gold catalyst, and with i- and n-shell growth times of 60 min and 30 min, respectively. The feeding ratios of silicon:boron and silicon:phosphorus are 500:1 and 200:1, respectively. **c**, High-resolution TEM image (spherical-aberration-corrected) of the p-i-n coaxial silicon nanowire. Scale bar, 5 nm.

[1]Department of Chemistry and Chemical Biology, [2]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.
*These authors contributed equally to this work.

diameters of ~300 nm (280–360 nm for other nanowires), which is in agreement with independent transmission electron microscopy (TEM) and atomic force microscopy measurements. In addition, high-resolution TEM images (Fig. 1c) confirm that the nanowire shell is indeed polycrystalline. We note that this nanocrystalline shell structure could enhance light absorption in the nanowires (see below).

To characterize electrical transport through the p-i-n coaxial silicon nanowires, we fabricated metal contacts selectively to the inner p-core and outer n-shell (Fig. 2a). Briefly, core/shell silicon nanowires were etched selectively using potassium hydroxide (KOH) solution (see Methods) to expose the p-core in a lithographically defined region, and then metal contacts were made to the p-core and n-shell after a second lithographic patterning step, as shown in the SEM images of Fig. 2b. Dark current–voltage (I–V) curves obtained from devices fabricated in this way (Fig. 2c) exhibit several notable features. First, the linear I–V curves from core–core (p1-p2) and shell–shell (n1-n2) configurations indicate that ohmic contacts are made to both core and shell portions of the nanowires. Second, the I–V curve for the shell–shell contact reveals a shell

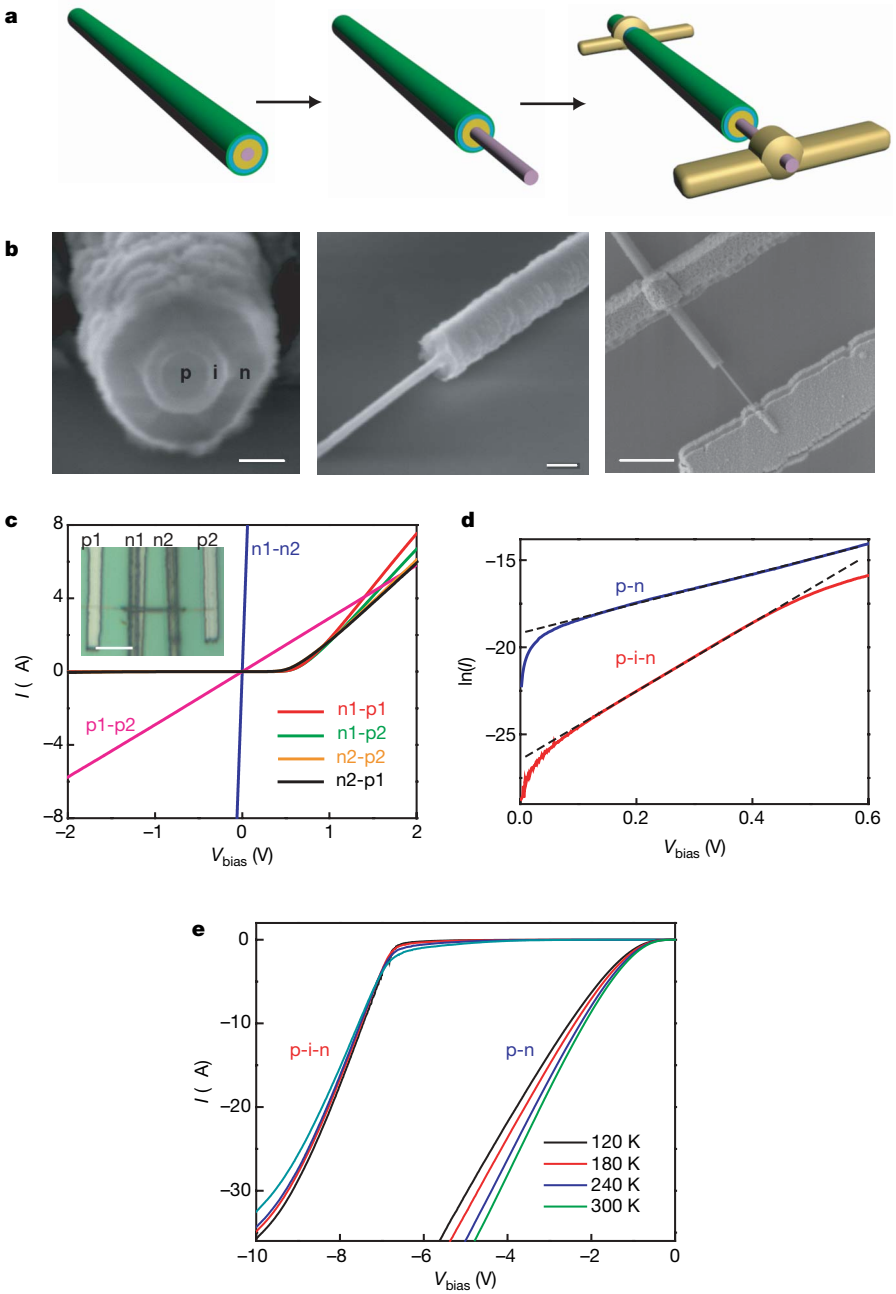


Figure 2 | Device fabrication and diode characterization. **a**, Schematics of device fabrication. Left, pink, yellow, cyan and green layers correspond to the p-core, i-shell, n-shell and PECVD-coated SiO_2, respectively. Middle, selective etching to expose the p-core. Right, metal contacts deposited on the p-core and n-shell. **b**, SEM images corresponding to schematics in **a**. Scale bars are 100 nm (left), 200 nm (middle) and 1.5 μm (right). **c**, Dark I–V curves of a p-i-n device with contacts on core–core, shell–shell and different core–shell combinations. V_{bias}, the applied bias voltage. Inset, optical microscope image of the device. Scale bar, 5 μm. **d**, Semi-log scale I–V curves of p-i-n and p-n diodes. The ideality factor N can be extrapolated (dashed lines) from the diode linear regimes (p-i-n diode, 0.12–0.50 V; p-n diode, 0.10–0.60 V), which are 1.96 and 4.52 for p-i-n and p-n diodes, respectively. To keep the total diameters of the p-n and p-i-n silicon nanowires approximately the same, the p-n silicon nanowire was grown with a 100-nm diameter gold catalyst and an n-shell growth time of 100 min. The SiH_4/dopants feeding ratios for p-n and p-i-n nanowires are the same (silicon:boron, 500:1; silicon:phosphorus, 200:1). **e**, Temperature-dependent I–V measurement of the p-n and p-i-n diodes in the reverse bias voltage regime.

conductance of 132 μS, higher than that of the core (3 μS); the calculated shell resistivity is within a factor of two of that measured for single-crystal n-type silicon nanowire prepared with a similar SiH_4:PH_3 ratio[15]. The highly conductive n-shell will reduce or eliminate potential drop along the shell, thereby enabling uniform radial carrier separation and collection when illuminated[8]. Third, *I*–*V* curves recorded from different core–shell contact geometries show rectifying behaviour, and demonstrate that the p-i-n coaxial silicon nanowires behave as well-defined diodes. The reproducibility of the selective etching and contact formation to p-cores and n-shells was further demonstrated by defining more complex 'AND' and 'OR' diode logic gates using single p-i-n coaxial silicon nanowires (Supplementary Fig. 1).

The core/shell silicon nanowire diodes were further characterized by analysing data recorded with and without the i-layer as a function of temperature. Fits to ln(*I*)–*V* data recorded in forward bias from p-i-n and p-n coaxial structures (Fig. 2d) are linear, and yield diode ideality factors *N* of 1.96 and 4.52, respectively (see Methods). The *N*-values show that introduction of the i-layer yields much better quality diodes. Reverse bias measurements from p-i-n and p-n diodes (Fig. 2e) also show markedly different behaviour: the p-i-n diode breaks down at much larger reverse-bias voltage (approximately −7 V) than the p-n diode (approximately −1 V) for all temperatures studied. In addition, the reverse-bias breakdown voltage of the p-n diode increases with decreasing temperature, which is consistent with a Zener (tunnelling) breakdown mechanism, whereas the breakdown voltage of the p-i-n structures exhibits little temperature dependence, suggesting contributions from tunnelling and avalanche mechanisms[16]. Overall, these results indicate that tunnelling or leakage currents are more significant in the p-n diode[17], and that the diode quality factor and breakdown behaviour are readily controlled during nanowire growth by the introduction of the i-layer as in planar structures[18,19].

The photovoltaic properties of the p-i-n coaxial silicon nanowire diodes were characterized under air mass 1.5 global (AM 1.5G) illumination. *I*–*V* data recorded from one of the better devices (Fig. 3a) yields an open-circuit voltage V_{oc} of 0.260 V, a short-circuit current I_{sc} of 0.503 nA and a fill factor F_{fill} of 55.0%. The maximum power output P_{max} for the silicon nanowire device at 1-sun (see Methods) is ~72 pW. Notably, these values were constant for measurements made over a seven-month period, thus demonstrating excellent stability of our nanowire photovoltaic elements. In addition, *I*–*V* data recorded using contacts to the n-shell that were 5.9 μm (n1) and 13.3 μm (n2) from the p-core contact (Fig. 3b) exhibited essentially the same photovoltaic response, thus indicating that the n-shell is equipotential with radial carrier separation occurring uniformly along the entire length of the core/shell silicon nanowire device. Measurements of I_{sc} as a function of the p-i-n coaxial silicon nanowire (Fig. 3c) length show linear scaling with values of 1 nA silicon nanowire^{-1} readily achieved for lengths of 10 μm (1-sun), whereas V_{oc} is essentially independent of length. The linear scaling of I_{sc} with silicon nanowire length suggests that photogenerated carriers are collected uniformly along the length of these radial nanostructures, and that scattering of light by the metal contacts does not make a major contribution to the observed photocurrent.

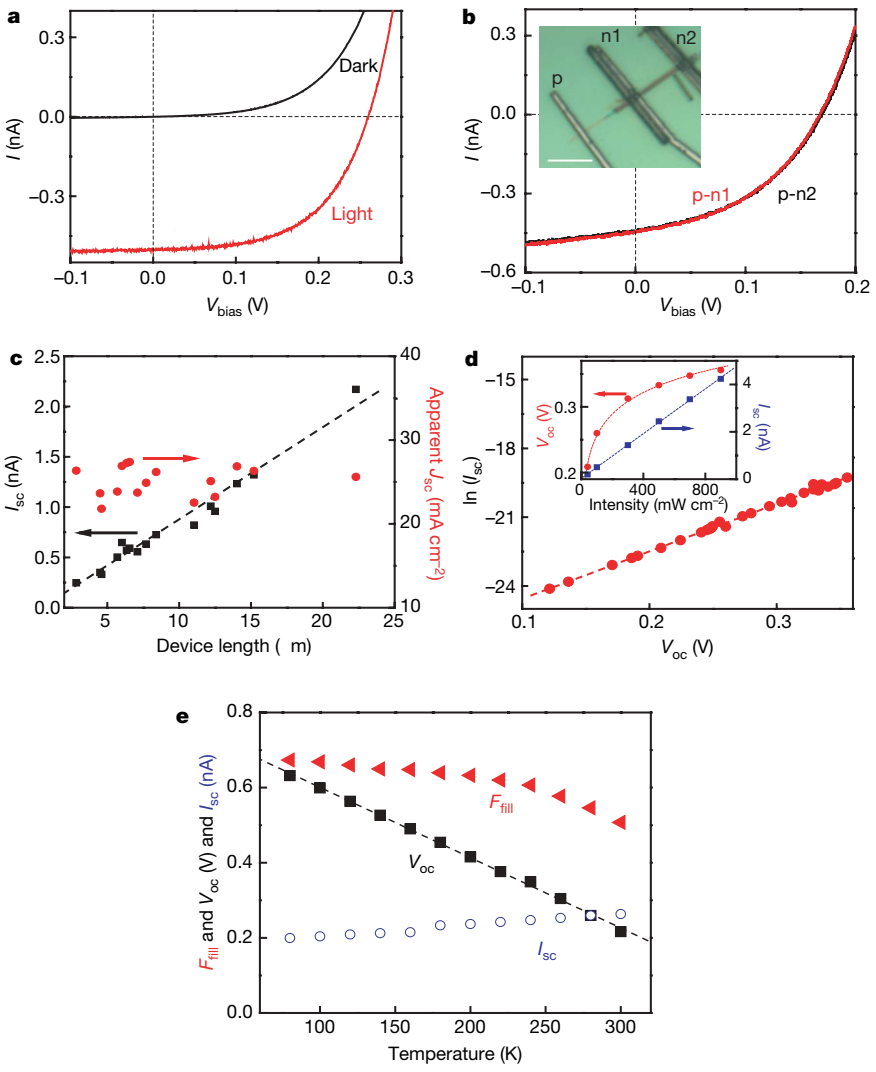


Figure 3 | Characterization of the p-i-n silicon nanowire photovoltaic device. **a**, Dark and light *I*–*V* curves. **b**, Light *I*–*V* curves for two different n-shell contact locations. Inset, optical microscopy image of the device. Scale bar, 5 μm. **c**, Device-length-dependent I_{sc} and J_{sc} (upper bound) plots. **d**, Plot of ln(I_{sc}) versus V_{oc}; each point corresponds to a different light intensity. Inset, light-intensity-dependent I_{sc} and V_{oc} plots. **e**, Temperature-dependent measurement. The device was illuminated at 0.6-sun to reduce sample heating, which may cause temperature fluctuations. The red triangle, black square and blue circle correspond to F_{fill}, V_{oc} and I_{sc}, respectively. The same device was characterized in **a**, **d** and **e**. The p-i-n coaxial silicon nanowire was grown using conditions as in Fig. 2.

The apparent short-circuit current density J_{sc} calculated using the projected area of the core/shell nanowire structure was $23.9 \pm 1.2\ mA\ cm^{-2}$ (upper bound, excluding metal covered and exposed p-core areas) and $16.0 \pm 0.8\ mA\ cm^{-2}$ (lower bound, including metal-covered and exposed p-core areas) for the device in Fig. 3a. The use of projected area to estimate the apparent current density is consistent with the methodology used with other nanostructured photovoltaic devices[3–6] and our use of devices as nanoscale power sources (see below, Fig. 4). Control experiments were also carried out to investigate the principal area of light absorption by devices. For example, measurements made on devices with and without lithographic masks that block illumination of the nanowire (Supplementary Fig. 2), with and without external scattering centres, and as a function of incident angle of illumination verify that the reported photocurrents and large apparent photocurrent densities arise primarily from direct nanowire absorption and are not much enhanced by scattering and/or waveguiding of incident light remote from devices. We note that the large nanowire J_{sc} values (Fig. 3c) imply substantial absorption across the solar spectrum. Such absorption is consistent with the nanocrystalline shell structure of the nanowires and previous studies of microcrystalline thin films[18], although the detailed nature of absorption will require further investigation. The apparent photovoltaic efficiency η of this device is $3.4 \pm 0.2\%$ (upper bound) and $2.3 \pm 0.2\%$ (lower bound), but might be improved through increased understanding of absorption and better coupling of light into the devices, for example, by vertical integration[8] or multilayer stacking[20].

I_{sc} and V_{oc} depend linearly and logarithmically, respectively, on the light intensity incident on the chip (inset, Fig. 3d), consistent with systematic increase in photogenerated carriers[19,21]. We note that the apparent efficiency is substantially higher at multiple-sun illumination: $4.1 \pm 0.2\%$ and $4.5 \pm 0.3\%$ (upper bounds) under 3-sun and 5-sun conditions, respectively. Although this apparent efficiency enhancement is larger than that in a planar silicon solar cell[19], it is consistent with the larger ideality factor (N) and lower 1-sun V_{oc} of the nanowire devices[19]. Analysis of a plot of $\ln(I_{sc})$ versus V_{oc} (Fig. 3d) yields values of the diode ideality factor and saturation current of $N = 1.86$ and $I_0 = 2.72$ pA, respectively (see Methods). These values are similar to those extrapolated from the dark measurements ($N = 1.96$, $I_0 = 3.24$ pA), and thus demonstrate good consistency in the behaviour and analysis of these core/shell silicon nanowire diode devices.

In addition, the temperature dependences of I_{sc}, V_{oc} and F_{fill} were characterized to understand better the behaviour of the silicon nanowire photovoltaic devices (Fig. 3e). I_{sc} decreases slightly with decreasing temperature, and can be attributed to reduced light absorption due to increasing bandgap as temperature is reduced[22]. V_{oc} exhibits a substantial linear increase with decreasing temperature, where the slope (dV_{oc}/dT) of $-1.9\ mV\ K^{-1}$ is close to the value ($-1.7\ mV\ K^{-1}$) calculated in single crystalline silicon solar cells[21]. The observed increase in V_{oc} can be attributed to a reduced recombination rate at lower temperature[21,22], and yields an apparent efficiency of 6.6% (upper bound) at 80 K (0.6-sun). The F_{fill} also increases with decreasing temperature (as expected from the negative dV_{oc}/dT)[22]. Taken together, these V_{oc} and F_{fill} results indicate that the silicon nanowire photovoltaic performance at room temperature (298 K in our experiments) can be significantly improved by reducing recombination processes, for example, by improving the crystalline structure of the shells and/or passivating the nanowire surface and grain boundaries[19,23].

Our core/shell silicon nanowire results can be compared to nanocrystal-based[4] and nanorod-based[5,6] photovoltaic devices. The best silicon nanowire device exhibits large apparent short-circuit current densities—$23.9\ mA\ cm^{-2}$ (upper bound) and $16.0\ mA\ cm^{-2}$ (lower bound)—with upper limits that are comparable to the $24.4\ mA\ cm^{2}$ value for the best thin film nanocrystalline silicon solar cell[24], and substantially better than values reported for CdSe nanorod/poly-3-hexathiophene[4] and dye-sensitized ZnO nanorod[5,6] solar cells. The V_{oc} value, 0.260 V, is 2–2.8 times lower than reported in these previous studies[4–6,24] and represents an area that should be addressed in future studies. However, the overall apparent efficiency of the p-i-n coaxial silicon nanowire photovoltaic elements—3.4% (upper bound) and 2.3% (lower bound)—exceeds reported nanorod/polymer and nanorod/dye systems[4–6], and could be increased substantially with improvements in V_{oc} by means of, for example, surface passivation. In addition, increasing the illumination intensity can yield stable improvements in the apparent efficiency of our p-i-n coaxial silicon nanowire photovoltaic elements in contrast to other nanostructured solar cells, which often exhibit degradation[4–6].

The ability of individual core/shell silicon nanowires to function as robust photovoltaic elements might indicate their potential as nanoscale power sources that might be integrated 'on-chip' with other

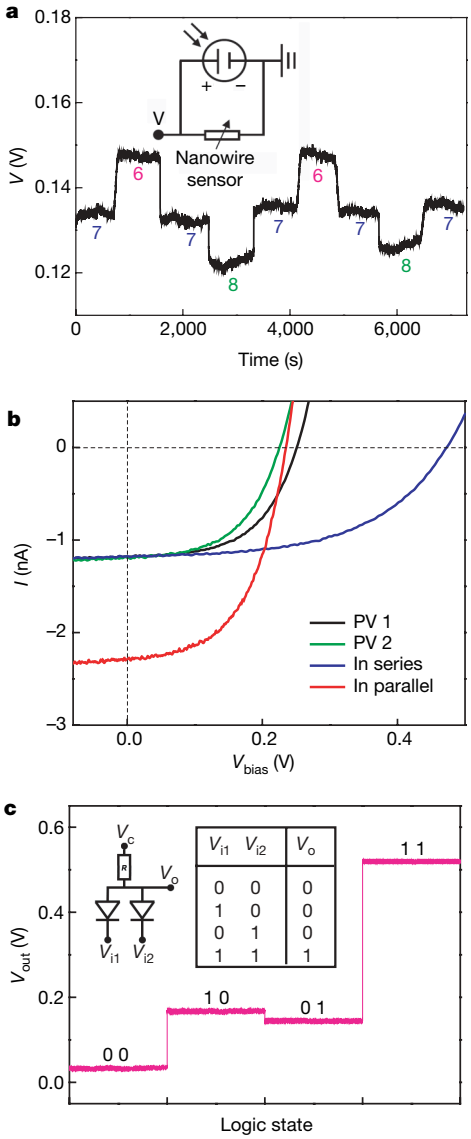


Figure 4 | Self-powered nanosystems. a, Real-time detection of the voltage drop across an aminopropyltriethoxysilane-modified silicon nanowire at different pH values. The silicon nanowire pH sensor is powered by a single silicon nanowire photovoltaic device operating under 8-sun illumination ($V_{oc} = 0.34$ V, $I_{sc} = 8.75$ nA). Inset, circuit schematics. **b**, Light I–V curves (1-sun, AM 1.5G) of two silicon nanowire photovoltaic devices (PV 1 and PV 2) individually and connected in series and in parallel. **c**, Nanowire AND logic gate powered by two silicon nanowire photovoltaic devices in series. Insets, circuit schematics and truth table for the AND gate. The resistance of CdSe nanowire is ~5 GΩ; the V_{oc} of two photovoltaic devices in series is 0.53 V. The large resistance of the CdSe nanowire and reverse-biased p-i-n diode makes V_c and V_i (HIGH) very close to V_{oc} of the photovoltaic device. To get V_i (LOW), the diode is simply grounded.

semiconductor nanowire- and carbon-nanotube-based nanoelectronic elements, given that these elements require power as low as a few nanowatts[25–27]. Recent work addressing this key issue has involved the use of piezoelectric ZnO nanowires for mechanical-to-electrical conversion, although the direct current (d.c.) power developed by this nanogenerator[28,29], 1–4 fW per nanowire, is at present less than is needed to drive nanoelectronic devices. Silicon nanowire photovoltaic elements can produce 50–200 pW per nanowire at 1-sun illumination, and thus could function as nanoscale power supplies for nanoelectronics by either increasing the light intensity or using several coupled elements. For example, a single silicon nanowire photovoltaic device, operating under 8-sun illumination (P_{max} = 1.86 nW, η = 4.8%) was used to drive a silicon nanowire pH sensor[25] without additional power (Fig. 4a). Measurements of the voltage drop across the p-type silicon nanowire sensor (powered solely by the silicon nanowire photovoltaic element) as a function of time (Fig. 4a) show reversible increase (or decrease) in voltage as the solution pH is decreased (or increased) that are consistent with the expected changes in resistance of the silicon nanowire with surface charge[25]. In addition, we note that the photovoltaic (under constant 8-sun illumination) and sensor devices both exhibited excellent stability over the approximately two-hour time of experiments.

Last, the core/shell silicon nanowire photovoltaic devices were interconnected in series and in parallel to demonstrate scaling of the output characteristics and to drive larger loads. *I–V* data recorded from two illuminated silicon nanowire elements (Fig. 4b) show several important features. First, the individual elements exhibit very similar behaviour, highlighting the good reproducibility of our core/shell nanowire devices. Second, interconnection of the two elements in series and parallel yields V_{oc} and I_{sc} values, respectively, that are approximately the sum of two, as expected. Notably, we have used interconnected silicon nanowire photovoltaic elements as the sole power supply driving a nanowire-based AND logic gate (Fig. 4c), where V_c and the voltage inputs 1 and 2 V_{i1} (HIGH) and V_{i2} (HIGH) are provided by two nanowire photovoltaic devices in series at 2-sun illumination. (HIGH is the input state and V_{i1} (HIGH) and V_{i2} (HIGH) are close to V_{oc} of the PV devices.) A summary of the input/output results (right inset, Fig. 4c) shows correct AND logic. This work thus demonstrates the potential for self-powered nanowire-based logic circuits and, more generally, the possibility of self-powered functional nanoelectronic systems through, for example, the integration of multiple stacked silicon nanowire photovoltaic elements with nanoelectronic, photonic and biological sensing devices.

METHODS SUMMARY

Single-crystalline silicon nanowire p-cores were synthesized by means of a nanocluster-catalysed VLS method[14,15], and then chemical vapour deposition was used to deposit i- and n-type nanocrystalline silicon shells; shell growth was carried out at higher temperature and lower pressure than those used in core growth to inhibit axial elongation of the silicon nanowire core. After nanowire growth, SiO_2 was deposited conformally by means of plasma-enhanced chemical vapour deposition (PECVD). Standard electron beam lithography, silicon wet chemical etching (KOH etchant) and thermal evaporation were used to make coaxial nanowire devices, with selective contacts on the p-core and n-shell. A standard solar simulator (150 W, Newport Stratford) with an AM 1.5G filter was used to characterize the photovoltaic device response, where the average intensity was calibrated using a power meter. For multiple-sun illumination, an aspheric lens was placed between the light source and nanowire devices. All electrical measurements were made with a probe station (TTP-4, Desert Cryogenics). For self-powered pH sensing and AND logic gate experiments, a computer-controlled analogue-to-digital converter (6030E, National Instruments) was used to record the voltage drop or voltage output of the silicon nanowire devices.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 May; accepted 7 August 2007.

1. Lewis, N. S. Toward cost-effective solar energy use. *Science* **315,** 798–801 (2007).
2. Lewis, N. S. & Crabtree, G. (eds) *Basic Research Needs for Solar Energy Utilization.* (Report of the Basic Energy Sciences Workshop on Solar Energy Utilization, US Department of Energy, Washington DC, 2005); <http://www.er.doe.gov/bes/reports/abstracts.html#SEU> (18–21 April, 2005).
3. Gratzel, M. Photoelectrochemical cells. *Nature* **414,** 338–344 (2001).
4. Huynh, W. U., Dittmer, J. J. & Alivisatos, A. P. Hybrid nanorod-polymer solar cells. *Science* **295,** 2425–2427 (2002).
5. Law, M., Greene, L. E., Johnson, J. C., Saykally, R. & Yang, P. Nanowire dye-sensitized solar cells. *Nature Mater.* **4,** 455–459 (2005).
6. Baxter, J. B. & Aydil, E. S. Nanowire-based dye-sensitized solar cells. *Appl. Phys. Lett.* **86,** 053114 (2005).
7. Luque, A., Marti, A. & Nozik, A. J. Solar cells based on quantum dots: multiple exciton generation and intermediate bands. *MRS Bull.* **32,** 236–241 (2007).
8. Kayes, B. M., Atwater, H. A. & Lewis, N. S. Comparison of the device physics principles of planar and radial p-n junction nanorod solar cells. *J. Appl. Phys.* **97,** 114302 (2005).
9. Zhang, Y., Wang, L. W. & Mascarenhas, A. Quantum coaxial cables. *Nano Lett.* **7,** 1264–1269 (2007).
10. Gust, D., Moore, T. A. & Moore, A. L. Mimicking photosynthetic solar energy transduction. *Acc. Chem. Res.* **34,** 40–48 (2001).
11. Klauk, H., Zschieschang, U., Pflaum, J. & Halik, M. Ultralow-power organic complementary circuits. *Nature* **445,** 745–748 (2007).
12. Browne, W. R. & Feringa, B. L. Making molecular machines work. *Nature Nanotechnol.* **1,** 25–35 (2006).
13. Avouris, P. & Chen, J. Nanotube electronics and optoelectronics. *Mater. Today* **9,** 46–54 (2006).
14. Wagner, R. S. & Ellis, W. C. Vapor–liquid–solid mechanism of single crystal growth. *Appl. Phys. Lett.* **4,** 89 (1964).
15. Zheng, G. F., Lu, W., Jin, S. & Lieber, C. M. Synthesis and fabrication of high-performance n-type silicon nanowire transistors. *Adv. Mater.* **16,** 1890–1893 (2004).
16. Hayden, O., Agarwal, R. & Lieber, C. M. Nanoscale avalanche photodiodes for highly sensitive and spatially resolved photon detection. *Nature Mater.* **5,** 352–356 (2006).
17. Karpov, V. G., Cooray, M. L. C. & Shvydka, D. Physics of ultrathin photovoltaics. *Appl. Phys. Lett.* **89,** 163518 (2006).
18. Shah, A. V. *et al.* Thin-film silicon solar cell technology. *Prog. Photovolt. Res. Appl.* **12,** 113–142 (2004).
19. Luque, A. & Hegedus, S. *Handbook of Photovoltaic Science and Engineering* (Wiley, Chichester, 2003).
20. Javey, A., Nam, S., Friedman, R. S., Yan, H. & Lieber, C. M. Layer-by-layer assembly of nanowires for three-dimensional, multifunctional electronics. *Nano Lett.* **7,** 773–777 (2007).
21. Würfel, P. *Physics of Solar Cells, From Principles to New Concepts* (Wiley-VCH, Weinheim, 2005).
22. Green, M. A. General temperature dependence of solar cell performance and implications for device modeling. *Prog. Photovolt. Res. Appl.* **11,** 333–340 (2003).
23. Aberle, A. G. Surface passivation of crystalline silicon solar cells: a review. *Prog. Photovolt. Res. Appl.* **8,** 473–487 (2000).
24. Green, M. A., Emery, K., King, D. L., Hishikawa, Y. & Warta, W. Solar cell efficiency tables (version 29). *Photovolt. Res. Appl.* **15,** 35–40 (2007).
25. Cui, Y., Wei, Q. Q., Park, H. K. & Lieber, C. M. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. *Science* **293,** 1289–1292 (2001).
26. Huang, Y. *et al.* Logic gates and computation from assembled nanowire building blocks. *Science* **294,** 1313–1317 (2001).
27. Bachtold, A., Hadley, P., Nakanishi, T. & Dekker, C. Logic circuits with carbon nanotube transistors. *Science* **294,** 1317–1320 (2001).
28. Wang, Z. L. & Song, J. Piezoelectric nanogenerators based on zinc oxide nanowire arrays. *Science* **312,** 242–246 (2006).
29. Wang, X., Song, J., Liu, J. & Wang, Z. L. Direct-current nanogenerator driven by ultrasonic waves. *Science* **316,** 102–105 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. W. Pang, D. C. Bell, H. G. Park, H. S. Choe, H. Yan and P. Xie for help with experiment and data analysis. C.M.L. acknowledges support from the MITRE Corporation and the Air Force Office of Scientific Research, and T.J.K. acknowledges an NSF graduate fellowship.

Author Contributions C.M.L., B.T., X.Z. and T.J.K. designed the experiments. B.T., X.Z., T.J.K., Y.F., N.Y. and G.Y. performed experiments and analyses. C.M.L., B.T., X.Z. and T.J.K. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.M.L. (cml@cmliris.harvard.edu).

METHODS

Nanowire synthesis. p-i-n coaxial silicon nanowires were prepared using 100-nm gold nanoclusters as catalysts, silane (SiH_4) as the silicon reactant, diboron (B_2H_6, 100 p.p.m. in H_2) as the p-type dopant, phosphine (PH_3, 1,000 p.p.m. in H_2) as the n-type dopant, and hydrogen (H_2) as the carrier gas. For the p-core nanowire growth, the flow rates of SiH_4, B_2H_6 and H_2 were 1, 10 and 60 standard (converted to standard temperature and pressure) cubic centimetres per minute, respectively. For the i-shell deposition, the flow rates of SiH_4 and H_2 were 0.15 and 60 standard cubic centimetres per minute, respectively, and 0.75 standard cubic centimetres per minute of PH_3 was added during the subsequent n-shell deposition. The growth temperatures for core and shells were 440 °C and 650 °C, respectively; the total pressures were 40 torr and 25 torr, respectively. The p-core growth lasted 3 h, and the deposition of i- and n- shells took 1 h and 0.5 h, respectively. Following the growth, the nanowire growth substrate was cleaned by oxygen plasma and the SiO_2 hard mask (30–60-nm thick) was deposited conformally onto the silicon nanowire surface by means of PECVD.

Device fabrication. All p-i-n devices were fabricated on heavily doped silicon substrates with 100 nm thermal oxide and 200 nm silicon nitride (n-type, resistivity $<0.005\,\Omega$ cm, Nova Electronic Materials). To fix the nanowires, chromium pads were patterned by electron beam lithography (EBL) and deposited directly on the SiO_2 hard mask by thermal evaporation. The second EBL step defined an etching window to expose the p-core in selected regions. The SiO_2 at the exposed nanowire region was first etched away using buffered HF with the e-beam resist as an etching mask, and the underlying shells of the silicon nanowire were further removed by KOH etching (70 °C, 45 s). In the last step, titanium/palladium contacts (3 nm/500 nm thick) at the p-core and n-shell of individual silicon nanowires were patterned by EBL and deposited by thermal evaporation. No annealing was required to ensure ohmic contact formation.

Sample illumination. A standard solar simulator (150 W, Newport Stratford) with an AM 1.5G filter was used in our experiments, in which the average intensity was calibrated using a power meter. For multiple-sun illumination, an aspheric lens was placed between the light source and the nanowire photovoltaic device. It should be noted that the exact light intensity incident on the nanowire cannot be measured exactly because its physical dimensions (around 300 nm in diameter and 3–22 μm in length) are orders of magnitude smaller than those of the pin hole of a power meter (millimetre range). Nevertheless, the intensity should be fairly close to 1-sun according to the light intensity uniformity guaranteed by the solar simulator vendor.

***I–V* data analysis.** The saturation current, I_0, together with the diode ideality factor N was extrapolated from the dark I–V curve using the ideal diode equation:

$$\ln(I) = \frac{q}{NkT} V + \ln(I_0)$$

where q is the electronic charge and k is the Boltzmann constant. The average $\pm 1\sigma$ ideality factor values obtained from the analysis of p-i-n and p-n coaxial silicon nanowire were 2.23 ± 0.13 and 4.70 ± 0.77, respectively. Under illumination, the ideal diode equation can be expressed in terms of I_{sc} and V_{oc} as:

$$\ln(I_{sc}) = \frac{q}{NkT} V_{oc} + \ln(I_0)$$

where fits to $\ln(I_{sc})$ versus V_{oc} (for example, Fig. 3d) are used to determine N and I_0 from the slope and intercept, respectively.

Silicon nanowire photovoltaic-powered nanowire sensor and logic devices. p-type silicon nanowire (diameter, 20 nm; Si:B = 16,000:1) sensor devices (Fig. 4a) were fabricated as described elsewhere[30]. The sensor devices were modified with aminopropyltriethoxysilane in ethanol/H_2O (95%/5%), and a single polydimethylsiloxane microfluidic channel was used to deliver different pH solutions during the experiments[30]. The silicon nanowire sensor resistance (10–30 MΩ) was chosen to allow for operation in the high-power working regime of the photovoltaic device in which the output voltage ranges from one-third to one-half of V_{oc} but the output current is relatively constant. The p-contact of the silicon nanowire photovoltaic device was connected to one end of the sensor device, whereas the n-contact and the other end of the sensor device were grounded, and a computer-controlled analogue-to-digital converter (6030E, National Instruments) was used to record the voltage drop across the silicon nanowire sensor. The self-powered AND gate (Fig. 4c) was made entirely from nanowires, in which p-i-n coaxial silicon nanowires were configured as the two diodes and a CdSe nanowire was used as the resistor. The large resistance of the CdSe nanowire and reverse-biased p-i-n diodes yielded V_c and V_i (HIGH) values close to the V_{oc} (0.53 V) of the photovoltaic device (two p-i-n coaxial silicon nanowire elements in series). Voltage outputs for all logic gate devices were recorded using a computer-controlled analogue-to-digital converter.

30. Patolsky, F., Zheng, G. F. & Lieber, C. M. Fabrication of silicon nanowire devices for ultrasensitive, label-free, real-time detection of biological and chemical species. *Nature Protocols* **1**, 1711–1724; doi:10.1038/nprot.2006.227 (2006).

LETTERS

Carbon dioxide release from the North Pacific abyss during the last deglaciation

Eric D. Galbraith[1]†, Samuel L. Jaccard[1,4], Thomas F. Pedersen[2], Daniel M. Sigman[3], Gerald H. Haug[4], Mea Cook[5], John R. Southon[6] & Roger Francois[1]

Atmospheric carbon dioxide concentrations were significantly lower during glacial periods than during intervening interglacial periods, but the mechanisms responsible for this difference remain uncertain. Many recent explanations call on greater carbon storage in a poorly ventilated deep ocean during glacial periods[1–5], but direct evidence regarding the ventilation and respired carbon content of the glacial deep ocean is sparse and often equivocal[6]. Here we present sedimentary geochemical records from sites spanning the deep subarctic Pacific that—together with previously published results[7]—show that a poorly ventilated water mass containing a high concentration of respired carbon dioxide occupied the North Pacific abyss during the Last Glacial Maximum. Despite an inferred increase in deep Southern Ocean ventilation during the first step of the deglaciation (18,000–15,000 years ago)[4,8], we find no evidence for improved ventilation in the abyssal subarctic Pacific until a rapid transition ~14,600 years ago: this change was accompanied by an acceleration of export production from the surface waters above but only a small increase in atmospheric carbon dioxide concentration[8]. We speculate that these changes were mechanistically linked to a roughly coeval increase in deep water formation in the North Atlantic[9–11], which flushed respired carbon dioxide from northern abyssal waters, but also increased the supply of nutrients to the upper ocean, leading to greater carbon dioxide sequestration at mid-depths and stalling the rise of atmospheric carbon dioxide concentrations. Our findings are qualitatively consistent with hypotheses invoking a deglacial flushing of respired carbon dioxide from an isolated, deep ocean reservoir[1–5,12], but suggest that the reservoir may have been released in stages, as vigorous deep water ventilation switched between North Atlantic and Southern Ocean source regions.

The rate at which a portion of the ocean interior exchanges gases with the atmosphere ('ventilation') is reflected by the concentrations of dissolved ^{14}C and O_2. Both are replenished by exchange with the atmosphere, but whereas ^{14}C decays at a globally uniform rate, O_2 consumption occurs only where organic matter is respired (Fig. 1). Thus, taken together, these tracers provide complementary information on ocean ventilation and accumulated organic matter respiration, giving key insights into ocean circulation, CO_2 sequestration and nutrient distribution.

The measurement of past variations in the ^{14}C content of subsurface water masses has long been sought, but has often proven difficult, particularly in waters of the deep Pacific. In theory, the surface-to-deep gradient of $\Delta^{14}C$ (see Fig. 1 legend) can be reconstructed simply by comparing the $^{14}C/^{12}C$ ratio measured in the carbonate tests of planktonic foraminifera with that of coeval benthic foraminifera. This approach has provided evidence for reduced ventilation in deep waters of both the Atlantic[9] and Southern[13] Oceans, and has shown variable patterns in the upper ocean[4,14]. However, the broad range among analogous measurements in the deep equatorial Pacific has prompted the suggestion that the ^{14}C activity of the deep North Pacific during the Last Glacial Maximum (LGM) was no different from that of today[6,15]. We tested this hypothesis by picking foraminifera in a core raised from 3.6 km water depth in the Gulf of Alaska (Supplementary Information).

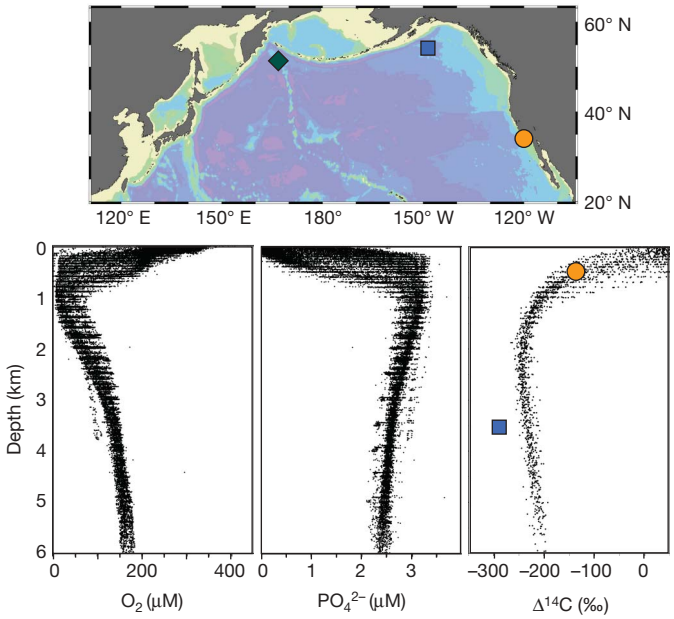


Figure 1 | Dissolved oxygen, phosphate and radiocarbon in the present-day North Pacific. Bottom panel, all measurements available in GLODAP[16] north of 20° N in the Pacific Ocean are shown (see Supplementary Information for additional discussion). $\Delta^{14}C$ is the $^{14}C/^{12}C$ of DIC expressed as the deviation from the $^{14}C/^{12}C$ of a standard (see Methods) in ‰. Top panel, locations of core sites GGC-37 (50.42° N, 167.72° E, 3,300 m) and ODP 882 (50.35° N, 167.58° E, 3,244 m) are shown by the dark green diamond, ODP 887 (54.37° N, 148.45° W, 3,647 m) by the dark blue square and ODP 893 by the dark yellow circle. The reconstructed $\Delta^{14}C'_{cont\text{-}atm}$ (Supplementary Information) of bottom waters at Sites 887 and 893 during the LGM are shown as the dark blue square and dark yellow circle, respectively, at their approximate palaeo-depths (120 m less than today).

[1]Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [2]School of Earth and Ocean Sciences, University of Victoria, Victoria, British Columbia V8W 3P6, Canada. [3]Department of Geosciences, Princeton University, Princeton, New Jersey 08544, USA. [4]Geological Institute, Department of Earth Sciences, ETH Zürich, Zürich CH-8092, Switzerland. [5]Department of Ocean Sciences, University of California, Santa Cruz, California 95064, USA. [6]Department of Earth System Science, University of California, Irvine, California 92697, USA. †Present address: Department of Atmospheric and Oceanic Sciences, Princeton University, Princeton, New Jersey 08544, USA.

Our results, plotted in Fig. 2, show an apparent age difference between co-occurring planktonic and benthic foraminifera during the LGM (18–20 kyr ago) of 1,600–1,900 uncorrected ^{14}C years; this is clearly greater than the apparent age difference of ~1,250 ^{14}C years near Site 887 today[16]. The present results indicate a greater LGM bottom water age than the low-latitude Pacific measurements from 2–2.8 km depth[6], but are indistinguishable from the planktonic–benthic difference estimated on a core from 3.2 km water depth in the eastern equatorial Pacific (Supplementary Information)[17]. Comparison to recently developed reconstructions of atmospheric ^{14}C (Supplementary Information) implies that the ^{14}C activity of the North Pacific at 3.6 km depth, relative to the contemporary atmosphere ($\Delta^{14}C'_{cont\text{-}atm}$, see Methods), was equivalent to about −290‰, a decrease of ~60‰ from the modern value. This indicates that the deep LGM North Pacific was substantially more isolated than the coeval deep Atlantic Ocean, which was, itself, relatively poorly ventilated ($\Delta^{14}C'_{cont\text{-}atm} > -215$‰, ref. 12). We note that benthic foraminifera measured at 2.7 km water depth on the New Zealand margin[13] imply a value of $\Delta^{14}C'_{cont\text{-}atm}$ of < -350‰, consistent with suggestions that deep Southern Ocean waters were more poorly ventilated still[4].

In contrast to the lower deep ocean, the ventilation of the upper North Pacific during the LGM seems to have been similar to or better than that of today, as previously shown by ^{14}C measurements on the California margin[14] (Figs 1, 2). This observation is consistent with a relatively vigorously circulating, vertically expanded equivalent of the North Pacific Intermediate Water within the upper 2 km of the water column[7,14]. Together with our results, this shows that the LGM vertical Δ^{14}C gradient between intermediate and abyssal waters of the North Pacific was stronger (>100‰) than that of today (~60‰), supporting the hypothesis that the isolation of LGM abyssal waters involved reduced vertical exchange with the intermediate waters above[2,5]. We note that this comparison is made only over the short time window of the LGM, as we do not have evidence to test whether this was a persistent feature throughout the glacial period.

Geochemical measurements of LGM sediments from two deep water locations in the subarctic Pacific (Fig. 1) complement our radiocarbon data in characterizing the isolated deep glacial water mass, showing that it bore lower oxygen concentrations and therefore harboured higher concentrations of respired CO_2. Sediments deposited during the LGM at both locations are enriched in authigenic U (Fig. 3c), which is known to precipitate in oxygen-depleted sediments. Similar glacial U accumulation has been observed at many sites in the deep sea[1,18,19], but its interpretation has often been confounded by the possibility that higher organic fluxes caused reducing conditions within the sediments themselves during the LGM, which could theoretically have overwhelmed any effect of bottom water oxygenation. Here, the export production proxies Ba/Al and opal, measured in the same sediments (Fig. 3g, h), show generally reduced fluxes during the last ice age (Supplementary Information), so that the enhanced U accumulation must have resulted from substantially lower bottom water O_2 (ref. 1). This requires the LGM North Pacific at 3–3.6 km depth to have contained a higher concentration of respired carbon than it does today. Although the absence of Mo enrichment (not shown) argues that anoxia did not develop in the vicinity of our core sites, oxygen concentrations may have dropped far below their current concentrations of ~130 μM during the LGM.

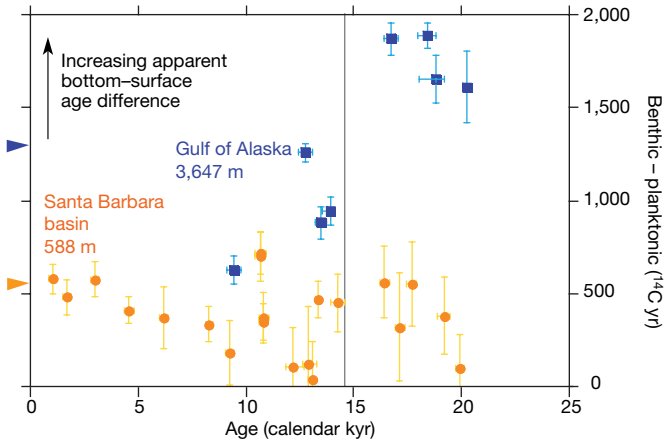


Figure 2 | Apparent age differences between paired benthic and planktonic foraminifera at intermediate and deep sites in the North Pacific over the past 25 kyr. Published data from the upper ocean are shown by dark yellow circles, using the published age model[14]. New data from ODP Site 887 are shown as dark blue squares. Differences are calculated from the raw, measured ^{14}C ages. Estimated pre-industrial values for both sites are indicated by the triangles at the left-hand side. The calendar ages used here are the calibrated planktonic ages. Errors are $\pm 1\sigma$ and include errors in the ^{14}C measurement, calendar age calculation and reservoir age. The vertical grey line is drawn at 14.6 kyr ago, as in Fig. 3.

It is widely thought that the state of the glacial deep ocean was fundamentally determined by Southern Ocean processes[1–5], so that the poor ventilation of the deep LGM Pacific would have been related to sea surface conditions near Antarctica. Ice core records show that deglacial warming of the Southern Ocean and a large rise in CO_2 partial pressure (p_{CO_2}) occurred between 18 and 15 kyr ago[8]. During this same period, ^{13}C (ref. 20) and ^{14}C (ref. 4) in the eastern equatorial Pacific thermocline both became progressively depleted, while the atmospheric ^{14}C activity dropped precipitously[15], suggesting that the p_{CO_2} increase was related to the exposure of poorly ventilated deep waters at the surface. Given this evidence, one might expect gradual signs of increased ventilation in the abyssal North Pacific to have accompanied the early southern warming, starting at 18 kyr ago. Surprisingly, the geochemical characteristics of the abyssal subarctic Pacific sediments register no hint of improved ventilation until an abrupt transition about 3 kyr later, witnessed by a host of proxies in the three cores (Fig. 3).

Calcium carbonate ($CaCO_3$), which was nearly or completely absent in these cores throughout much of the glacial sediment, suddenly appears in abundance at ~14.6 kyr ago at each of the three deep subarctic Pacific sites (Fig. 3d, Supplementary Fig. 3). Although increases in local $CaCO_3$ export may have strongly affected sedimentary $CaCO_3$ concentrations, the magnitude of change suggests substantially reduced dissolution at the sea floor, implying higher bottom water CO_3^{2-} concentration owing to decreased dissolved inorganic carbon (DIC) and/or increased alkalinity[21]. This implication mirrors high deglacial CO_3^{2-} concentrations previously inferred from foraminiferal Zn/Ca measurements in the deep equatorial Pacific[22]. An abrupt >0.3‰ increase of δ^{13}C in the epibenthic foraminifer *Cibicidoides* over the same sediment interval (Fig. 3e)[7] is consistent with a rapid reduction of the respired DIC concentration (Supplementary Information), suggesting that this was an important contributor to accelerated $CaCO_3$ burial.

Over the same sedimentary interval, the δ^{18}O of planktonic foraminifera reveals an abrupt freshening and/or warming of surface waters (Fig. 3i)[7]. Within the error of our ^{14}C-based age models (that is, a few centuries, see Supplementary Information) this change in sea surface conditions occurred during the sudden warming in the North Atlantic at the start of the Bølling period (Fig. 3j). Although centennial-scale phasing between the two basins cannot be resolved, the near-synchrony suggests a strong dynamical link. Furthermore, records of ^{14}C within the North Atlantic interior that span the deglaciation[9,10] describe a pattern of changing ventilation that is remarkably similar to that described here. Like our reconstruction from the North Pacific, Atlantic records show a strengthened vertical ^{14}C gradient during the LGM, with relatively well-ventilated waters overlying extremely ^{14}C-depleted abyssal waters, a situation that clearly persisted in the Atlantic throughout the early CO_2 rise from 18 to 15 kyr ago. This was followed, ~14.6 kyr ago, by a rapid ventilation of

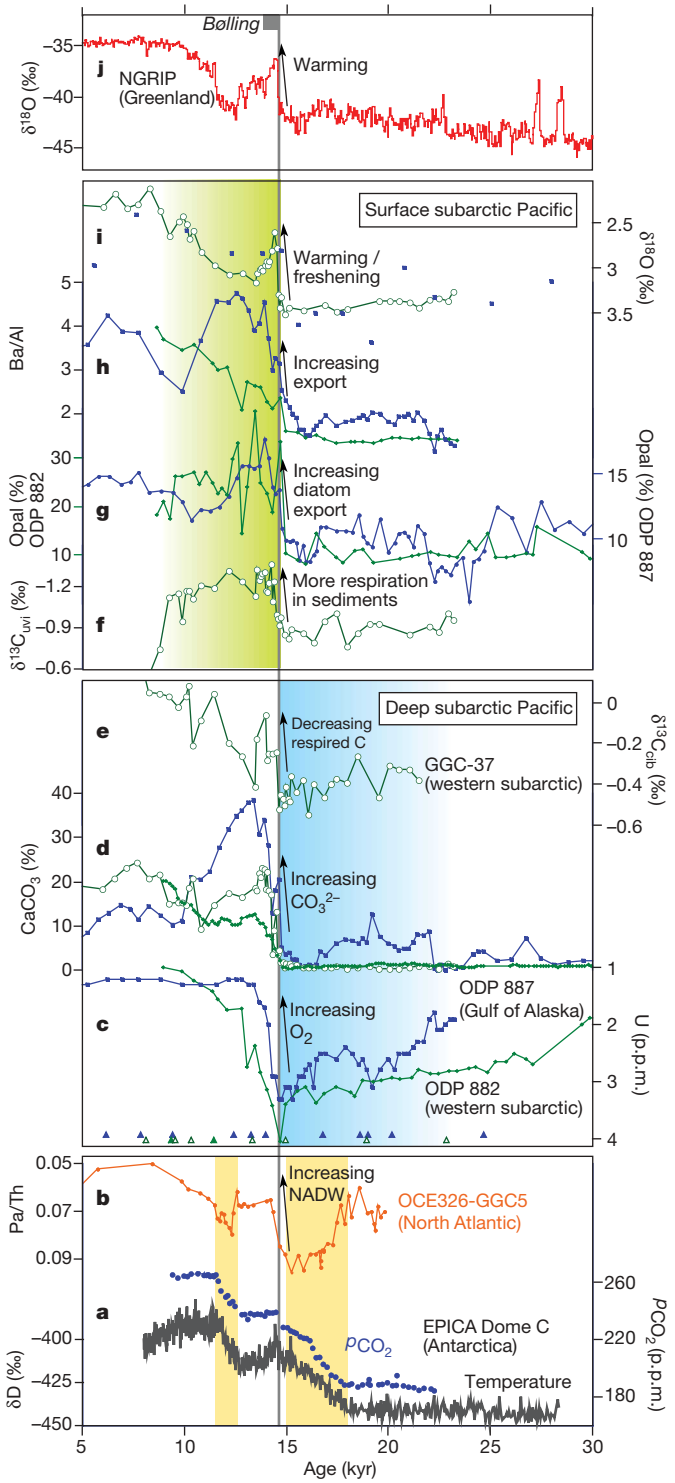


Figure 3 | Multi-proxy sedimentary records from the subarctic Pacific spanning 5 to 30 kyr ago. New measurements are shown from sites ODP 887 (dark blue squares) and ODP 882 (small dark green diamonds), with previously published data from GGC-37 (open green circles, ref. 7; see Fig. 1 for locations). Proxies reflect the chemical characteristics of bottom waters (**c**–**e**), surface water fertility (**f**–**h**) and surface temperature/salinity (**i**). Small triangles show calibrated planktonic ^{14}C dates for ODP 887 (dark blue, filled), ODP 882 (dark green, filled) and GGC-37 (dark green, open). Also shown are published records of Greenland temperature (**j**, $\delta^{18}O$, ref. 33), NADW formation (**b**, Pa/Th, ref. 11) and Antarctic temperature (δD) and CO_2 (**a**, ref. 8, using the timescale of ref. 4). Note that foraminiferal $\delta^{13}C$ and $\delta^{18}O$ records from GGC-37 include additional measurements that were not presented in the original publication (L. Keigwin, personal communication). $\delta^{13}C = [(^{13}C/^{12}C)_{sample}/(^{13}C/^{12}C)_{standard} - 1] \times 1{,}000‰$; $\delta^{18}O = [(^{18}O/^{16}O)_{sample}/(^{18}O/^{16}O)_{standard} - 1] \times 1{,}000‰$; the reported values correspond to the Pee Dee Belemnite standard. The blue and green shaded areas highlight the periods before and after 14.6 kyr ago, respectively. The yellow shaded areas indicate the deglacial periods of Antarctic warming, when p_{CO_2} was increasing.

deep waters in both basins, when vigorous formation of North Atlantic Deep Water (NADW) resumed[9–11]. The apparent inter-basin synchrony and chemical homogeneity of the modern abyssal Pacific both suggest that this mid-deglacial ventilation typified a large fraction of the global abyss; the North Pacific and North Atlantic, combined, account for more than 36% of the global ocean volume.

However, although the mid-deglacial ventilation of northern deep waters is synchronous, within chronological uncertainty, with an abrupt ~10 p.p.m. increase of p_{CO_2} recorded in Antarctic ice[8] (Fig. 3a), this represents only a small fraction of the total deglacial increase of p_{CO_2}, the majority of which occurred during two phases of Antarctic warming (Fig. 3a). It is perhaps counterintuitive that increased ventilation of such a large deep ocean volume would fail to cause a more sustained rise in atmospheric CO_2 about 14.6 kyr ago. However, sinking NADW bears low concentrations of unused nutrients, and thus its reinvigoration should have been associated with enhanced net global storage of respired carbon in the ocean interior[23–25]. We suggest that this occurred through an increase of respired DIC concentrations at intermediate depths, compensating the observed ventilation of deep northern oceans.

Our export production proxy data (Fig. 3) suggest that at least part of this compensatory intermediate-depth carbon storage took place in the North Pacific. Across the same narrow sedimentary intervals that witnessed rapid improvement of deep ocean ventilation, Ba/Al ratios and opal concentrations rise sharply on both sides of the subarctic Pacific (Fig. 3g, h). Although a decrease in the flux of lithogenic material also occurred at this time, ^{230}Th-normalized flux measurements (Supplementary Information) confirm an acceleration in the rain of organic detritus[21] and siliceous plankton to the sea floor. Further support for a deglacial increase of export production comes from $\delta^{13}C$ measurements of the infaunal foraminifera *Uvigerina* (Fig. 3f, Supplementary Information)[7] and from parallel observations throughout the subarctic[26,27] and northeast Pacific[28,29], indicating a basin-wide increase in the rate at which nutrients were supplied to the surface ocean. The apparent coincidence of increased export production in the North Pacific with NADW rejuvenation is also consistent with recent findings from a global ocean-ecosystem model[30], whereby greater Atlantic overturning increases the upward flux of remineralized nutrients from the abyss to the global surface ocean. In short, when NADW formation is enhanced, the global thermocline shoals, facilitating the upwelling and entrainment of nutrient-bearing deep waters into the wind-driven circulation.

We therefore suggest that the entrainment of North Pacific deep waters into the wind-driven circulation increased dramatically near the start of the Bølling, accompanied by an enhanced influx of deep waters from the south, establishing a more 'estuarine' circulation in the North Pacific. The upward flux of nutrients supported the widespread boom in export production during the Bølling[26–29], and the resulting rain of organic matter drove a depletion of oxygen within the upper ocean, explaining the coeval intensification of the intermediate-depth oxygen-minimum zone previously described[31]. Meanwhile, the intense oxygen depletion reflects enhanced storage of respired DIC in the thermocline that counteracted the removal of DIC from the abyssal waters below. Although this minimized the short-term effect on CO_2, we note that the removal of respired DIC from the deep ocean would have caused the CO_3^{2-} activity there to increase[12], deepening the lysocline; such deepening has long been recognized as a potential mechanism to deplete the oceanic alkalinity inventory and, hence, decrease global CO_2 solubility[12,22]. Thus, the removal of respired DIC from a large fraction of the deep sea would

have caused an additional, long-term increase of p_{CO_2} (ref. 32), helping to propel the climate system into the interglacial period.

METHODS SUMMARY

Benthic and planktonic foraminifera were hand picked from samples that each spanned a vertical interval of <2 cm, extracted at broad benthic foraminifera abundance peaks (Supplementary Information). Calendar-year ages were calculated assuming a constant subarctic Pacific reservoir age of $\Delta R = 550 \pm 250$ yr (Supplementary Information). Biogenic opal concentrations were determined by molybdate-blue spectrophotometry on alkaline extracts. $CaCO_3$ concentrations were quantified by coulometric CO_2 determinations, assuming no other carbonate-bearing phase was present. Absolute elemental concentrations of homogenized powders were measured by inductively coupled plasma-mass spectrometry (ICP-MS) for ODP 882, following acid digestion, and by inductively coupled plasma-optical emission spectrometry (ICP-OES) and ICP-MS for ODP 887, following fusion and subsequent dissolution in acid. The age models for ODP 887 and GGC-37 are based exclusively on calibrated planktonic ^{14}C measurements over the deglaciation, whereas that of ODP 882 is also tied to that of the neighbouring core GGC-37 at the midpoint of the rapid $CaCO_3$ rise (Supplementary Information).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 December 2006; accepted 7 September 2007.

1. Francois, R. *et al.* Contribution of Southern Ocean surface-water stratification to low atmospheric CO_2 concentrations during the last glacial period. *Nature* **389,** 929–935 (1997).
2. Toggweiler, J. R. Variation of atmospheric CO_2 by ventilation of the ocean's deepest water. *Paleoceanography* **14,** 571–588 (1999).
3. Stephens, B. B. & Keeling, R. F. The influence of Antarctic sea ice on glacial–interglacial CO_2 variations. *Nature* **404,** 171–174 (2000).
4. Marchitto, T. *et al.* Marine radiocarbon evidence for the mechanism of deglacial atmospheric CO_2 rise. *Science* **316,** 1456–1459 (2007).
5. Sigman, D. M. & Boyle, E. A. Glacial/interglacial variations in atmospheric carbon dioxide. *Nature* **407,** 859–869 (2000).
6. Broecker, W. *et al.* Ventilation of the glacial deep Pacific Ocean. *Science* **306,** 1169–1172 (2004).
7. Keigwin, L. D. Glacial-age hydrography of the far northwest Pacific Ocean. *Paleoceanography* **13,** 323–339 (1998).
8. Monnin, E. *et al.* Atmospheric CO_2 concentrations over the last glacial termination. *Science* **291,** 112–114 (2001).
9. Robinson, L. F. *et al.* Radiocarbon variability in the western North Atlantic during the last deglaciation. *Science* **310,** 1469–1473 (2005).
10. Skinner, L. C. & Shackleton, N. J. Rapid transient changes in northeast Atlantic deep water ventilation age across Termination I. *Paleoceanography* **19,** doi:10.1029/2003PA000983 (2004).
11. McManus, J. F. *et al.* Collapse and rapid resumption of Atlantic meridional circulation linked to deglacial climate changes. *Nature* **428,** 834–837 (2004).
12. Boyle, E. A. Vertical oceanic nutrient fractionation and glacial/interglacial CO_2 cycles. *Nature* **331,** 55–56 (1988).
13. Sikes, E. L., Samson, C. R., Guilderson, T. P. & Howard, W. R. Old radiocarbon ages in the southwest Pacific Ocean during the last glacial period and deglaciation. *Nature* **405,** 555–559 (2000).
14. Kennett, J. P. & Ingram, B. L. A 20,000-year record of ocean circulation and climate change from the Santa Barbara basin. *Nature* **377,** 510–514 (1995).
15. Broecker, W. & Barker, S. A 190‰ drop in atmosphere's $\Delta^{14}C$ during the "Mystery Interval" (17.5 to 14.5 kyr). *Earth Planet. Sci. Lett.* **256,** 90–99 (2007).
16. Key, R. M. *et al.* A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP). *Glob. Biogeochem. Cycles* **18,** doi:10.1029/2004GB002247 (2004).
17. Shackleton, N. J. *et al.* Radiocarbon age of last glacial Pacific deep water. *Nature* **335,** 708–711 (1988).
18. Dezileau, L., Bareille, G. & Reyss, J.-L. Enrichissement en uranium authigène dans les sédiments glaciaires de l'océan Austral. *CR Geosci.* **334,** 1039–1046 (2002).
19. Sarkar, A., Bhattacharya, S. K. & Sarin, M. M. Geochemical evidence for anoxic water in the Arabian Sea during the last glaciation. *Geochim. Cosmochim. Acta* **51,** 1009–1016 (1993).
20. Spero, H. J. & Lea, D. W. The cause of carbon isotope minimum events on glacial terminations. *Science* **296,** 522–525 (2002).
21. Jaccard, S. L. *et al.* Glacial/interglacial changes in subarctic North Pacific stratification. *Science* **308,** 1003–1006 (2005).
22. Marchitto, T. M., Lynch-Stieglitz, J. & Hemming, S. R. Deep Pacific $CaCO_3$ compensation and glacial–interglacial atmospheric CO_2. *Earth Planet. Sci. Lett.* **231,** 317–336 (2005).
23. Sigman, D. M. & Haug, G. H. in *Treatise on Geochemistry* Vol. 6 (eds Holland, D. & Turekian, K. K.) 491–528 (Elsevier, Amsterdam, 2003).
24. Ito, T. & Follows, M. J. Preformed phosphate, soft tissue pump and atmospheric CO_2. *J. Mar. Res.* **63,** 813–839 (2005).
25. Toggweiler, J. R. *et al.* Representation of the carbon cycle in box models and GCMs — 2. Organic pump. *Glob. Biogeochem. Cycles* **17,** doi:10.1029/2001GB001841 (2003).
26. Crusius, J. *et al.* Influence of northwest Pacific productivity on North Pacific Intermediate Water oxygen concentrations during the Bolling-Allerod interval (14.7–12.9 ka). *Geology* **32,** 633–636 (2004).
27. Brunelle, B. G. *et al.* Evidence from diatom-bound nitrogen isotopes for Subarctic Pacific stratification during the last ice age and a link to North Pacific denitrification changes. *Paleoceanography* **22,** doi:10.1029/2005PA001205 (2007).
28. Ivanochko, T. S. & Pedersen, T. F. Determining the influences of Late Quaternary ventilation and productivity variations on Santa Barbara Basin sedimentary oxygenation: a multi-proxy approach. *Quat. Sci. Rev.* **23,** 467–480 (2004).
29. Ortiz, J. D. *et al.* Enhanced marine productivity off western North America during warm climate intervals of the past 52 k.y. *Geology* **32,** 521–524 (2004).
30. Schmittner, A. Decline of the marine ecosystem caused by a reduction in the Atlantic overturning circulation. *Nature* **434,** 628–633 (2005).
31. Zheng, Y. *et al.* Intensification of the northeast Pacific oxygen minimum zone during the Bolling-Allerod warm period. *Paleoceanography* **15,** 528–536 (2000).
32. Broecker, W. & Peng, T. H. The role of $CaCO_3$ compensation in the glacial to interglacial atmospheric CO_2 change. *Glob. Biochem. Cycles* **1,** 15–29 (1987).
33. Andersen, K. K. *et al.* High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature* **431,** 147–151 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. de Vernal, J. Leduc, P. Dulski, M. Soon and K. Gordon for analytical assistance, and J. Sarmiento, R. Toggweiler, M. Kienast, L. Keigwin and S. Calvert for intellectual and practical support. R. Schlitzer's program Ocean Data View was used to generate Fig. 1. E.D.G., T.F.P. and R.F. were supported by the Natural Sciences and Engineering Research Council of Canada and the Canadian Foundation for Climate and Atmospheric Sciences, S.L.J. by a Swiss National Foundation post-doctoral fellowship, D.M.S. by US NSF, and by BP and Ford Motor Company through the Princeton Carbon Mitigation Initiative, and G.H.H. by Deutsche Forschungsgemeinschaft.

Author Contributions E.D.G. and S.L.J. contributed equally to this work. T.F.P. and G.H.H. initiated and guided the project. E.D.G. prepared samples and picked foraminifera from ODP Site 887, S.L.J. prepared and analysed samples from site ODP Site 882. R.F. and S.L.J. made the ^{230}Th measurements and J.R.S. made the radiocarbon measurements. M.C. contributed to the ^{14}C analysis. E.D.G., S.L.J. and D.M.S. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.D.G. (egalbrai@princeton.edu).

METHODS

Radiocarbon measurements. Radiocarbon was measured at the UC Irvine Keck AMS facility using the NIST OX1 radiocarbon standard as a reference. All results obtained for the LGM and deglaciation are reported in Supplementary Table 1 according to the conventions of ref. 34. Size-dependent backgrounds were determined using calcite blanks, and were checked by measurements of small aliquots of the IAEA C2 and FIRI turbidite secondary standards. Uncertainties are based on the scatter in repeated measurements as well as counting statistics, and contributions from normalization to the OX1 standard and from background subtraction are included. Ages of planktonic foraminifera were calibrated to calendar years using CALIB 5.0.2 (refs 35, 36).

For the calculations of LGM ^{14}C activity shown in Fig. 1, the $\Delta^{14}C$ of a given benthic foram sample at the time of calcification was reconstituted to the calibrated age of the coexisting planktonic foraminifera, given the ^{14}C decay rate, providing the palaeo-bottom water $\Delta^{14}C$ (relative to the modern atmospheric ^{14}C activity). Changes in atmospheric $\Delta^{14}C$ through time were provided by the IntCal04 compilation[37]. Comparisons between benthic and atmospheric $\Delta^{14}C$ were made by the two methods shown in Supplementary Fig. 2, to assess the importance of large secular changes in atmospheric $\Delta^{14}C$ on the timescales of deep Pacific ventilation (>1,500 yr)[38]. The raw arithmetic bottom-atmosphere differences calculated directly from these methods are deceiving, however, in that the values are typically reported in terms of their deviation from the relatively low ^{14}C activity of the modern atmosphere, in parts per thousand. To be directly comparable to modern oceanographic measurements of ^{14}C, such as those shown in Fig. 1, we propose a definition of the ^{14}C activity relative to the contemporary atmospheric ^{14}C activity:

$$\Delta^{14}C' = (\Delta^{14}C_{bot} - \Delta^{14}C_{atm})/(\Delta^{14}C_{atm} + 1{,}000) \times 1{,}000‰$$

where $\Delta^{14}C_{bot}$ is the reconstituted bottom water $\Delta^{14}C$, and $\Delta^{14}C_{atm}$ is the reference palaeo-atmospheric $\Delta^{14}C$. The values shown in Fig. 1 for ODP 887 and ODP 893 are the averages of all available measurements from 16.4 to 20.5 kyr ago, calculated both as $\Delta^{14}C'_{proj\text{-}atm}$ and $\Delta^{14}C'_{cont\text{-}atm}$ (Supplementary Fig. 2). For Site 887, these calculated averages were $-295 \pm 38‰$ and $-286 \pm 32‰$, respectively ($\pm 1\sigma$, $n = 4$), while for Site 893 they were $-153 \pm 36‰$ and $-137 \pm 35‰$, respectively ($n = 5$). An estimate of the minimum possible ventilation decrease at Site 887 can be made by assuming a very small LGM reservoir age of $\Delta R = 250$ yr, which gives $\Delta^{14}C_{proj\text{-}atm}$ and $\Delta^{14}C_{cont\text{-}atm}$ of $-272 \pm 28‰$ and $-257 \pm 19‰$, respectively.

Geochemical measurement methods. All analyses were made on freeze-dried and homogenized samples. Biogenic opal percentage was determined by alkaline extraction of silica[39]. Replicate measurements indicate a reproducibility of $\pm 3\%$. For ODP 882, absolute elemental concentrations of U, Ba and Al were measured by ICP-MS (ELAN 5000A) using solution nebulization after mixed acid digestion (HF-$HClO_4$) under pressure. Precision and accuracy were better than 5%. For ODP 887, samples were fused with $LiBO_2$ at 1,273 K and the resulting glass dissolved in HNO_3. Using aliquots of the same solutions, absolute elemental concentrations of Al were measured by ICP-OES (Varian Vista Pro) while U and Ba were measured by ICP-MS (ELAN 9000) by ALS Chemex Ltd. Accuracy was better than 5%, 5% and 2%, respectively, for replicate measurements. All $CaCO_3$ concentrations were measured by coulometry as described[21].

Age models. The age model for ODP 887 is based on 12 planktonic ^{14}C ages (Supplementary Table 2), calibrated as described above with $\Delta R = 550 \pm 250$ yr. Changes in reservoir age are likely to have rendered the median calibrated values inaccurate at most times, but the true values should tend to lie within the error limits, if these are accurately estimated. Therefore, we chose three values at the calibrated 1σ error limits and one at the calibrated 2σ error limit that minimize changes in sedimentation rate (see selected ages, Supplementary Table 2). The age model was linearly interpolated between the selected ages.

For GGC-37, the published ^{14}C ages of ref. 8 were recalibrated using Calib 5.0.2 with a reservoir age ΔR of 550 ± 250 yr. The suggestion that reservoir ages in this region changed markedly between 12.8 and 13.3 ^{14}C kyr BP (ref. 40) calls into question the reliability of ^{14}C ages within this interval. We therefore did not include ^{14}C ages within this interval, in favour of a constant sedimentation rate. This conservative approach suggests that the midpoint of the rapid $CaCO_3$ concentration rise was at ~14.5 kyr ago, compared to the age of 15.3 kyr ago suggested by the age model of ref. 7. The age model for ODP 882 is tied to the age model of the nearby core GGC-37 at the midpoint of the rapid $CaCO_3$ rise (Supplementary Information), supplemented with two calibrated planktonic radiocarbon ages (at 9.8 and 12.0 kyr ago), and linearly interpolated between age ties to coincide with the age model of ref. 21 at 30 kyr ago. Note that the age model for ODP 887 remains completely independent of the other records, and no attempt was made to improve the fit of the planktonic $\delta^{18}O$ records to the Greenland temperature record.

34. Stuiver, M. & Polach, H. A. Reporting of ^{14}C data. *Radiocarbon* **19,** 355–363 (1977).
35. Stuiver, M. & Reimer, P. J. Extended ^{14}C data base and revised CALIB 3.0 ^{14}C age calibration program. *Radiocarbon* **35,** 215–230 (1993).
36. Hughen, K. A. *et al.* Cariaco Basin calibration update: Revisions to calendar and ^{14}C chronologies for core PL07–58PC. *Radiocarbon* **46,** 1161–1187 (2004).
37. Reimer, P. J. *et al.* IntCal04 terrestrial radiocarbon age calibration, 0–26 cal kyr BP. *Radiocarbon* **46,** 1029–1058 (2004).
38. Adkins, J. F. & Boyle, E. A. Changing atmospheric $\Delta^{14}C$ and the record of deep water paleoventilation ages. *Paleoceanography* **12,** 337–344 (1997).
39. Mortlock, R. A. & Froelich, P. N. A simple method for the rapid-determination of biogenic opal in pelagic marine-sediments. *Deep-Sea Res. A* **36,** 1415–1426 (1989).
40. Sarnthein, M. *et al.* Warmings in the far northwestern Pacific promoted pre-Clovis immigration to America during Heinrich event 1. *Geology* **34,** 141–144 (2006).

LETTERS

The rapid drift of the Indian tectonic plate

Prakash Kumar[1], Xiaohui Yuan[2], M. Ravi Kumar[1], Rainer Kind[2,3], Xueqing Li[2] & R. K. Chadha[1]

The breakup of the supercontinent Gondwanaland into Africa, Antarctica, Australia and India about 140 million years ago, and consequently the opening of the Indian Ocean, is thought to have been caused by heating of the lithosphere from below by a large plume whose relicts are now the Marion, Kerguelen and Réunion plumes. Plate reconstructions based on palaeomagnetic data suggest that the Indian plate attained a very high speed (18–20 cm yr^{-1} during the late Cretaceous period) subsequent to its breakup from Gondwanaland, and then slowed to ~5 cm yr^{-1} after the continental collision with Asia ~50 Myr ago[1,2]. The Australian and African plates moved comparatively less distance and at much lower speeds of 2–4 cm yr^{-1} (refs 3–5). Antarctica remained almost stationary. This mobility makes India unique among the fragments of Gondwanaland. Here we propose that when the fragments of Gondwanaland were separated by the plume, the penetration of their lithospheric roots into the asthenosphere were important in determining their speed. We estimated the thickness of the lithospheric plates of the different fragments of Gondwanaland around the Indian Ocean by using the shear-wave receiver function technique. We found that the fragment of Gondwanaland with clearly the thinnest lithosphere is India. The lithospheric roots in South Africa, Australia and Antarctica are between 180 and 300 km deep, whereas the Indian lithosphere extends only about 100 km deep. We infer that the plume that partitioned Gondwanaland may have also melted the lower half of the Indian lithosphere, thus permitting faster motion due to ridge push or slab pull.

The term lithosphere, commonly understood as describing Earth's rigid outer shell floating on a viscous asthenosphere, originally evolved in a mechanical sense[6] to explain the post-glacial rebound phenomenon. Since then, several usages of this term have evolved, such as thermal, chemical and seismic lithospheres[7]. Traditionally, seismologists refer to this boundary as the Gutenberg discontinuity after the discovery of low-velocity zones in regions underlying oceanic basins by Gutenberg[8]. Old and stable continental regions are understood to be underlain by a thick lithosphere[9], as demonstrated by the presence of numerous diamondiferous regions located within their interiors. Results from seismic tomography show the presence of deep roots in old continents such as Africa where the lithospheric thickness exceeds 250 km (ref. 10).

Alterations of the primordial lithospheric configuration due to passage over hotspots in areas covered by vast regions of basalt on the continent (large igneous provinces) are also shown by thinning of the lithosphere and by the presence of low-velocity uppermost mantle. Because the thickness of the lithosphere has a prominent role in shielding the mantle attrition processes that are vital for determining the stability factor for the survival of the Precambrian crust, its precise determination is important. In addition, imprints of major tectonic events such as passage over hotspots (plume–lithosphere interaction), rifting due to continental breakup, and continental collision are expected to be manifested as alterations in the deep lithospheric architecture.

We apply a recently developed seismic method (shear-wave (S) receiver functions) to determine with high accuracy the depth of the lithosphere/asthenosphere boundary (LAB) in the region of the Indian Ocean and the surrounding fragments of Gondwanaland. Figure 1 and Supplementary Fig. 1 show the distribution of the seismic stations used. This method uses S-to-P converted waves from the LAB beneath a seismic station. Details of the technique and examples of applications in other regions have been given in several papers[11–18]. The observed S receiver functions are shown in Fig. 2a for each station. These data are summation traces of several tens of records at each station. Two prominent phases are clearly visible at all stations, marked Moho and LAB. To verify our observations of the LAB, we show in Supplementary Fig. 3 synthetic S receiver functions and their relation to possible anisotropy in the mantle and to compressional-wave (P) receiver function observations. Individual S receiver functions of HYB and three other Indian stations are shown in Supplementary Figures 4 and 5.

Conversions from the Moho and the LAB have different signs because they result from discontinuities with velocities that increase (Moho) and decrease (LAB) downwards. The times for the Moho vary between ~2 s and 8 s, and those for the LAB vary between about 4 s and 32 s. The Moho and LAB times, along with their corresponding depths (using the global reference model IASP91), are given in Supplementary Table 1. The stations in Fig. 2a are sorted in order of increasing LAB time. The depth of the LAB varies between 30 and 300 km. There is no obvious correlation between crustal thickness and LAB depth. Theoretical receiver functions are shown in Fig. 2b for the models in Fig. 2c. The agreement between computed and observed seismograms is very good. A simple model with a homogeneous crust and a homogeneous lithosphere above the asthenosphere can explain most features of the observations. Only the depths of the Moho and of the LAB need to be varied. In this study we have not attempted to model the sharpness of the discontinuities, but the relative amplitudes of the Moho and LAB phases with respect to their corresponding direct S phases clearly reveal that the amplitude at Moho is twofold to threefold that for LAB. An interesting feature of Fig. 2d is that the cratonization of the lithosphere is reflected in a decrease in the LAB amplitudes, whereas the Moho amplitudes remain nearly stable from oceanic regions to cratons.

For verification that a large LAB time corresponds to a thick high-velocity mantle lid, we examined the arrival times of the P-to-S converted waves from the discontinuity at 410 km, which roughly sample the average velocity above 410 km depth. The existence of a thick high-velocity lid can cause the converted waves to travel faster, at a rate proportional to the lid thickness. Consequently, the times of the P-to-S converted waves from the 410-km discontinuity should anticorrelate with the times of the S-to-P conversions from the LAB. In Fig. 3, we show the measured times from the LAB and from the 410-km and 660-km discontinuities. The waveform data of the conversions of the data from the 410-km and 660-km discontinuities are shown in Supplementary Fig. 6. Figure 3 shows clearly that the times from the LAB and the 410-km discontinuity are anticorrelated,

[1]National Geophysical Research Institute, Hyderabad 500 007, India. [2]GeoForschungsZentrum, 14473 Potsdam, Germany. [3]Freie Universität, Berlin 12249, Germany.

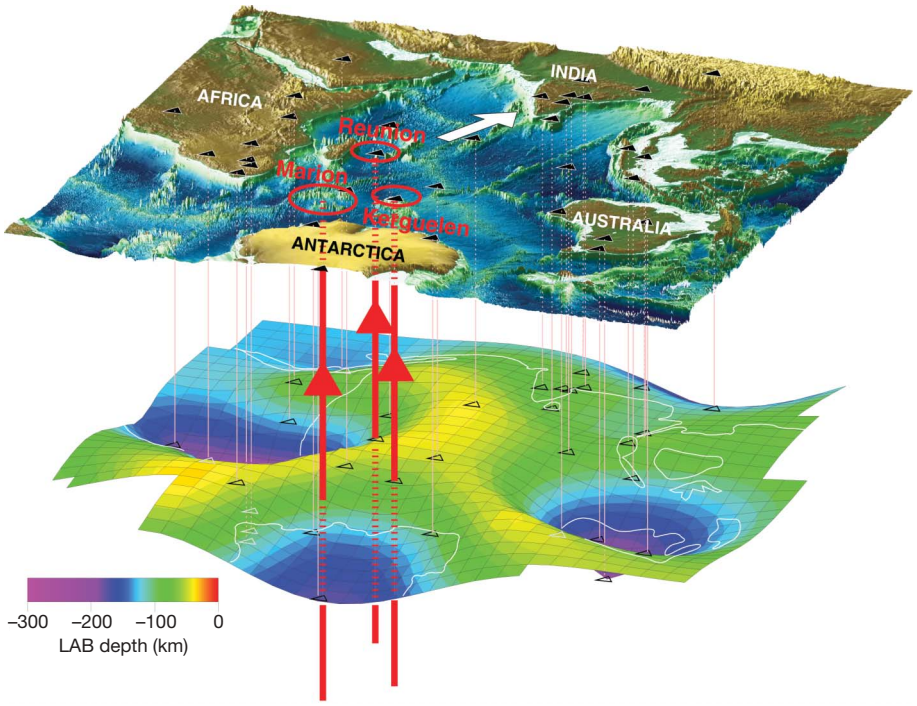


Figure 1 | Topography of the surface and the LAB in the region of the Indian Ocean and the fragments of Gondwanaland surrounding it. The Indian lithosphere is exceptionally thin compared with the other fragments of Gondwanaland. Black triangles denote seismic stations. The station locations are also shown in Supplementary Fig. 1, with station codes. Red circles mark the surface locations of the mantle plumes whose conduits are illustrated by the thick vertical lines.

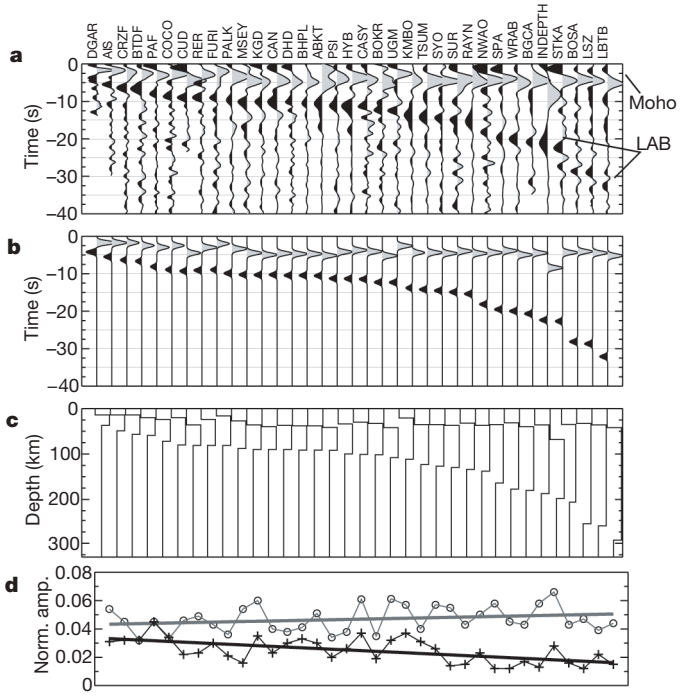


Figure 2 | S-receiver function data and modelling. a, Stacked S-receiver function traces from all stations used. Converted signals from the Moho and the LAB are clearly visible. The traces are arranged with increasing LAB time from left to right. The vertical axis measures the time differences between the S-to-P converted phases and the corresponding reference S phases. Time is negative because the S-to-P converted waves travel faster than the S waves. **b**, Synthetic seismograms[33], which model the observed waveforms very well. **c**, Simple models used for the computation of the theoretical seismograms. The models consist of a homogeneous crust, mantle lithosphere and asthenosphere. Velocities are fixed for all the models (V_s for the crust is 3.58 km s^{-1}, V_s for the lithospheric mantle lid is 4.5 km s^{-1} and V_s for the asthenosphere is 3.9 km s^{-1}). Only the depths of both discontinuities are varied to fit the travel times. **d**, Variation in normalized amplitudes (norm. amp.; with respect to the direct S phase) of the Moho (open circles) and the LAB (plus signs) with respect to the amplitude of the direct S phase. The grey and black lines are trend lines from a least-squares fit.

whereas those of the 410-km discontinuity and the 660-km discontinuity are correlated. Small LAB depths correlate with low average velocities above the 410-km discontinuity. The correlation of the times from the 410-km and 660-km discontinuities suggests that most of the time variations can be attributed to mantle velocity variation above 410 km depth. The possible influence of the topography of the 410-km discontinuity on this conclusion is small[19].

As shown in Fig. 1, on the basis of the data given in Fig. 2a and Supplementary Table 1, the lithosphere is very thin in the young regions of the mid-oceanic ridges and is very thick (more than 180 km) below the cratons of South Africa, Antarctica and Australia. The Indian lithosphere is ~100 km thick or less, although it was a part of the same Gondwanaland. Six stations on the Indian shield, namely DHD (situated on the western Dharwar craton), HYB and CUD (situated on the Eastern Dharwar craton), KGD (situated within the Godavari rift zone), and BHPL and BOKR in northern India, indicate a lithospheric thickness of 80–100 km. Such a thin lithosphere for the Indian shield is unexpected, because all these stations are sited on Archaean basement, away from the coast. Earlier studies of the lithospheric thickness in India do not lead to a homogeneous picture. Surface-wave tomography studies reveal that the lithosphere could be ~100 km (ref. 20) or 150 km (ref. 21) thick. Lithospheric thickness estimates based on temperature–depth profiles[22] yield an average thickness of 104 km for the Indian shield region. Studies based on temperature data constrained by pressure–temperature estimates from xenoliths lead to a thickness of 200–250 km (ref. 23) or 160 km (ref. 24) (adding S velocity data). Our estimates of lithospheric thickness are based purely on the seismic body-wave observations, yielding the present-day lithospheric thickness. We also consider that the resolution of our body-wave data is higher than that of the long-period surface-wave data. The lithosphere in the south Indian shield must originally have been thick, in view of the presence of diamond-bearing kimberlites[24] close to HYB. There is now an increasing use of structural and metamorphic pressure–temperature data, precise dating measurements and comparison of subsurface geophysical models to aid in palaeocontinent reconstructions with respect to Gondwanaland and Rodinia[25]. Diamonds originate in the deep roots of ancient continental blocks that extend into the diamond stability field beneath ~140 km and more. Ar–Ar dating of the kimberlites in the Indian shield constrains their age to the Proterozoic era (~1,091 ± 20 Myr (ref. 26)). Because

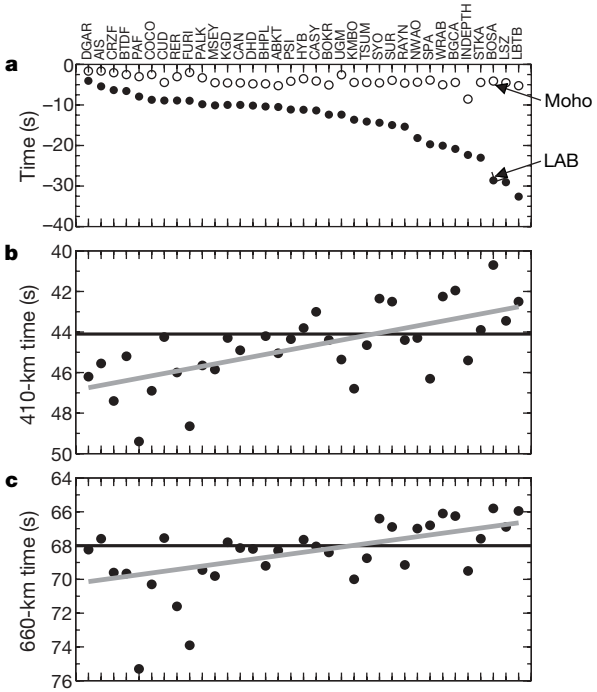


Figure 3 | Comparison of different upper mantle seismic phases. **a**, Filled circles show times for the S-to-P conversions at the LAB arranged in increasing order; open circles are those for the Moho. **b**, Times for the P-to-S converted waves at the 410-km discontinuity. **c**, Times for the P-to-S converted waves at the 660-km discontinuity. As the LAB times (**a**) increase, the corresponding times for the 410-km (**b**) and 660-km (**c**) discontinuities decrease. This observed anticorrelation confirms the interpretation of the LAB pulses as originating from the LAB. Scatterings in **b** and **c** are due to time picking error and to the presence of local heterogeneities within the upper mantle.

Figure 4 | Reconstruction of Permian Gondwanaland. The contours show the present-day continental lithospheric thicknesses.

the diamonds are older than the kimberlites that transport them to the surface, the lithosphere in this region must have been thicker before the breakup of Gondwanaland.

The high speed attained by India during the Cretaceous period coupled with the present-day estimates prompts us to argue that the originally thick lithosphere beneath India seems to have been preferentially thinned by a large plume either during or immediately after the Gondwana breakup ~130 Myr ago[2]. Subsequently, the Indian plate could have been further degenerated by the influence of its passage over the hotspots and the large-scale magmatic extrusions such as the Deccan and Rajmahal traps, although their role in thinning of the lithosphere cannot be ascertained. Osmium isotopic studies suggest a lack of evidence for the involvement of the subcontinental mantle lithosphere in Rajmahal basalts[27], which seem to share an origin with the Kerguelen lavas[28]. Figure 4 shows a reconstruction of Gondwanaland with the present-day lithospheric thickness, which is exceptionally small beneath India and east Africa. The loss of the lithospheric roots might have been the reason that the Indian plate attained a very high velocity of 20 cm yr^{-1}, which is unusual for any continental lithospheric plate[29]. The plate speed, resulting from either a push caused by a hot mantle source or a trench pull due to a cold downwelling slab, increases with decreasing root depth[30].

In the present study, the thickest lithosphere (~300 km) is found in the oldest continental nuclei in the Kaapvaal craton (Supplementary Fig. 2), which is diamond bearing, implying that the lithosphere is preserved over Archaean times. The other stations in the Archaean Kaapvaal craton also indicate a lithosphere thicker than 250 km. This result is supported by independent studies of seismic tomography[10]. As expected, the thickness of the lithosphere found by seismic tomography[31] is also greater than 200 km in the Australian and Antarctic shield regions. Tomographic inversion of teleseismic P and S travel times indicates that high-velocity lithosphere beneath the Tanzania craton extends to a depth of at least 200 km and possibly to 300 or 350 km (ref. 32). In addition, the lithosphere tends to get thinner in regions closer to the coast as a result of the effect of rifting due to the breakup of Gondwanaland.

Received 25 April; accepted 31 August 2007.

1. Klootwijk, C. T., Gee, J. S., Peirce, J. W. & Smith, G. M. An early India–Asia contact: Paleomagnetic constraints from Ninetyeast Ridge, ODP Leg 121. *Geology* **20,** 395–398 (1992).
2. Gaina, C., Müller, R. D., Brown, B. & Ishihara, T. Breakup and early seafloor spreading between India and Antarctica. *Geophys. J. Int.* **170,** 151–169 (2007).
3. Veevers, J. J. Breakup of Australia and Antarctica estimated as mid-Cretaceous (95 ± 5 Ma) from magnetic and seismic data at the continental margin. *Earth Planet. Sci. Lett.* **77,** 91–99 (1986).
4. Cande, S. C. & Mutter, J. C. A revised identification of the oldest sea-floor spreading anomalies between Australia and Antarctica. *Earth Planet. Sci. Lett.* **58,** 151–160 (1982).
5. Jacoby, W. R. in *Landolt-Bornstein, New Series, Group V* Vol. 2b (eds Fuchs, K. & Soffel, H.) 298–369 (Springer, Berlin, 1985).
6. Barrell, J. The strength of the Earth's crust. *J. Geol.* **22,** 655–683 (1914).
7. Anderson, D. L. Lithosphere, asthenosphere, and perisphere. *Rev. Geophys.* **33,** 125–149 (1995).
8. Gutenberg, B. *Physics of the Earth's Interior*. (Elsevier, New York, 1959).
9. Jordan, T. H. The continental tectosphere. *Rev. Geophys. Space Phys.* **13,** 1–13 (1975).
10. James, D. E., Fouch, M. J., VanDecar, J. C. & van der Lee, S. Tectospheric structure beneath southern Africa. *Geophys. Res. Lett.* **28,** 2485–2488 (2001).
11. Li, X., Kind, R., Yuan, X., Wölbern, I. & Hanka, W. Rejuvenation of the lithosphere by the Hawaiian plume. *Nature* **427,** 827–829 (2004).
12. Kumar, P. *et al.* The lithosphere–asthenosphere boundary in the North West Atlantic Region. *Earth Planet. Sci. Lett.* **236,** 249–257 (2005).
13. Kumar, P., Yuan, X., Kind, R. & Kosarev, G. The lithosphere–asthenosphere boundary in the Tien Shan-Karakoram region from S receiver functions—evidence of continental subduction. *Geophys. Res. Lett.* **32,** L07305, doi:10.1029/2004GL022291 (2005).
14. Kumar, P., Yuan, X., Kind, R. & Ni, J. Imaging the colliding Indian and Asian lithospheric plates beneath Tibet. *J. Geophys. Res.* 111, B06308 10.1029/2005JB003930 (2006).
15. Sodoudi, F., Yuan, X., Liu, Q., Kind, R. & Chen, J. Lithospheric thickness beneath the Dabie Shan, central eastern China from S receiver functions. *Geophys. J. Int.* **166,** 1363–1367 (2006).
16. Angus, D. A., Wilson, D. C., Sandvol, E. & Ni, J. F. Lithospheric structure of the Arabian and Eurasian collision zone in eastern Turkey from S-wave receiver functions. *Geophys. J. Int.* **166,** 1335–1346 (2006).
17. Landes, M., Ritter, J. R. R. & Readman, P. W. Proto-Iceland plume caused thinning of Irish lithosphere. *Earth Planet. Sci. Lett.* **255,** 32–40 (2007).

18. Li, X., Yuan, X. & Kind, R. The lithosphere–asthenosphere boundary beneath the western United States. *Geophys. J. Int.* **170,** 700–710 (2007).
19. Li, X. *et al.* Seismic observation of narrow plumes in the oceanic upper mantle. *Geophys. Res. Lett.* **30** (6), 1334, doi:10.1029/2002GL015411 (2003).
20. Polet, J. & Anderson, D. L. Depth extent of cratons as inferred from tomographic studies. *Geology* **23,** 205–208 (1995).
21. Mitra, S., Priestley, K., Gaur, V. K. & Rai, S. S. Shear-wave structure of the south Indian lithosphere from Rayleigh wave phase–velocity measurements. *Bull. Seism. Soc. Am.* **96,** 1551–1559 (2006).
22. Pandey, O. P. & Agarwal, P. K. Lithospheric mantle deformation beneath the Indian cratons. *J. Geol.* **107,** 683–692 (1999).
23. Artemieva, I. M. Global 1 degree × 1 degree thermal model TC1 for the continental lithosphere: Implications for lithosphere secular evolution. *Tectonophysics* **416,** 245–277 (2006).
24. Priestley, K. & McKenzie, D. The thermal structure of the lithosphere from shear velocities. *Earth Planet. Sci. Lett.* **244,** 285–301 (2006).
25. Mezger, K. & Cosca, M. A. The thermal history of the Eastern Ghats belt (India) as revealed by U–Pb and $^{40}Ar/^{39}Ar$ dating of metamorphic and magmatic minerals: implications for the SWEAT correlation. *Precambr. Res.* **94,** 251–271 (1999).
26. Kumar, A., Padma Kumari, V. M., Dayal, A. M., Murty, D. S. N. & Gopalan, K. Rb–Sr ages of Proterozoic Kimberlites of India: evidence for contemporaneous emplacement. *Precambr. Res.* **62,** 227–237 (1993).
27. Ingle, S., Scoates, J. S., Weis, D., Brügmann, G. & Kent, R. W. Origin of Cretaceous continental tholeiites in southwestern Australia and eastern India: insights from Hf and Os isotopes. *Chem. Geol.* **209,** 83–106 (2004).
28. Coffin, M. *et al.* Kerguelen hotspot magma output since 130 Ma. *J. Petrol.* **43,** 1121–1139 (2002).
29. Jurdy, D. M. & Gordon, R. G. Global plate motions relative to the hotspots 64 to 56 Ma. *J. Geophys. Res.* **89,** 9927–9936 (1984).
30. Gurnis, M. & Torsvik, T. H. Rapid drift of large continents during the late Precambrian and Paleozoic: Paleomagnetic constraints and dynamic models. *Geology* **22,** 1023–1026 (1994).
31. Gaherty, J. B., Mamoru, K. & Jordan, T. H. Seismological structure of the upper mantle: a regional comparison of seismic layering. *Phys. Earth Planet. Inter.* **110,** 21–41 (1999).
32. Ritsema, J., Nyblade, A. A., Owens, T. J., Langston, C. A. & VanDecar, J. C. Upper mantle seismic velocity structure beneath Tanzania, east Africa: Implications for the stability of cratonic lithosphere. *J. Geophys. Res.* **103** (B9), 21201–21214 (1998).
33. Frederiksen, A. W. & Bostock, M. G. Modelling teleseismic waves in dipping anisotropic structures. *Geophys. J. Int.* **141,** 401–412 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Phipps Morgan, L. Brown, D. Eaton and D. Mueller for helpful comments. This research was supported by the Deutsche Forschungsgemeinschaft. P.K. was supported by a DAAD Fellowship under which part of this work was performed at GeoForschungsZentrum. The director of the National Geophysical Research Institute is thanked for his support. Seismic data for most of the stations are available through the open data archives of IRIS (http://www.iris.edu/), GEOFON (http://www.gfz-potsdam.de/geofon/) and GEOSCOPE (http://geoscope.ipgp.jussieu.fr/), and Indian stations are supported by the Department of Science and Technology and IMD, India. Seismic data analysis was performed in SeismicHandler (K. Stammler).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.K. (kind@gfz-potsdam.de).

LETTERS

Anaerobic oxidation of short-chain hydrocarbons by marine sulphate-reducing bacteria

Olaf Kniemeyer[1]†, Florin Musat[1], Stefan M. Sievert[2], Katrin Knittel[1], Heinz Wilkes[3], Martin Blumenberg[4], Walter Michaelis[4], Arno Classen[5], Carsten Bolm[5], Samantha B. Joye[6] & Friedrich Widdel[1]

The short-chain hydrocarbons ethane, propane and butane are constituents of natural gas. They are usually assumed to be of thermochemical origin[1], but biological formation of ethane and propane has been also observed[2]. Microbial utilization of short-chain hydrocarbons has been shown in some aerobic species[3,4] but not in anaerobic species of bacteria. On the other hand, anaerobic utilization of short-chain hydrocarbons would in principle be expected because various anaerobic bacteria grow with higher homologues ($\geq C_6$)[5]. Indeed, chemical analyses of hydrocarbon-rich habitats with limited or no access of oxygen indicated *in situ* biodegradation of short-chain hydrocarbons[6–10]. Here we report the enrichment of sulphate-reducing bacteria (SRB) with such capacity from marine hydrocarbon seep areas. Propane or *n*-butane as the sole growth substrate led to sediment-free sulphate-reducing enrichment cultures growing at 12, 28 or 60 °C. With ethane, a slower enrichment with residual sediment was obtained at 12 °C. Isolation experiments resulted in a mesophilic pure culture (strain BuS5) that used only propane and *n*-butane (methane, isobutane, alcohols or carboxylic acids did not support growth). Complete hydrocarbon oxidation to CO_2 and the preferential oxidation of ^{12}C-enriched alkanes were observed with strain BuS5 and other cultures. Metabolites of propane included iso- and *n*-propylsuccinate, indicating a subterminal as well as an unprecedented terminal alkane activation with involvement of fumarate. According to 16S ribosomal RNA analyses, strain BuS5 affiliates with *Desulfosarcina/Desulfococcus*, a cluster of widespread marine SRB. An enrichment culture with propane growing at 60 °C was dominated by *Desulfotomaculum*-like SRB. Our results suggest that diverse SRB are able to thrive in seep areas and gas reservoirs on propane and butane, thus altering the gas composition and contributing to sulphide production.

Saturated hydrocarbons (alkanes) belong to the chemically least reactive organic compounds, but are nevertheless used by diverse microorganisms. Whereas biodegradation of alkanes with oxygen has been well-known for a century, their anaerobic biodegradation[5] still awaits a deeper understanding with respect to biochemical reactions and the range of utilizable chain lengths. An intensely investigated process in this respect is the anaerobic oxidation of methane with sulphate in marine sediments[5]. On the other hand, several anaerobic bacteria have been isolated that utilize saturated hydrocarbons with six or more carbon atoms[5]. Hence, there is a gap in our knowledge—are there microorganisms that can anaerobically degrade hydrocarbons with shorter chain lengths, in particular ethane, propane and *n*-butane? We therefore attempted to find representatives of sulphate-reducing microorganisms with such capacity by way of enrichment from marine hydrocarbon seeps. Enrichments were set up in defined anoxic media, using sediment from the Gulf of Mexico[11,12] for potentially cold-adapted SRB, and sediment from the Guaymas basin (Gulf of California)[13,14] for mesophilic and thermophilic SRB.

Slow, yet clearly hydrocarbon-dependent, sulphide production was observed in all incubations with propane or butane, and in an incubation with ethane at 12 °C (Table 1). The thermophilic enrichment culture with propane exhibited a substantial sulphide (>12 mM) production within three months, whereas other cultures with propane or butane reached similar concentrations within approximately six months. In subcultures, the thermophilic enrichment with propane and the mesophilic enrichment with butane attained the fastest growth. Ethane-dependent sulphate reduction remained very slow (Supplementary Fig. 1). Sulphate reduction with isobutane has not been observed thus far.

From the mesophilic enrichment with butane, a pure culture, strain BuS5, was isolated (Fig. 1a, left). Strain BuS5 and enrichment cultures with propane and butane were phylogenetically analysed on the basis of 16S rRNA gene sequences (Fig. 1b). Retrieved sequences were used to design 16S rRNA-specific fluorescent probes for visualization of abundant phylotypes in the enrichment cultures via cell hybridization (Supplementary Table 1; Fig. 1a; Supplementary Fig. 3). Strain BuS5 was a member of the *Desulfosarcina/Desulfococcus* cluster within the Deltaproteobacteria. This cluster comprises widespread bacteria, many of which have been detected in seep areas. Strain BuS5 was a major component (~50% of cells) of the parental mesophilic enrichment culture (Supplementary Fig. 3). A close relative of strain BuS5 was dominant (75% of cells) in the cold-adapted enrichment culture with butane (Butane12-GMe; Fig. 1). The thermophilic enrichment culture with propane (Propane60-GuB) was dominated (92% of cells) by a Gram-positive (*Desulfotomaculum*) species (Fig. 1a). The very slow sulphate-reduction with ethane (Supplementary Fig. 1) has not so far yielded a sediment-free subculture for comparable hybridization studies.

Strain BuS5 was characterized in more detail. Its growth occurred between 10 and 33 °C (optimum, 28 °C) and pH 6.0 and 7.4 (optimum, pH 6.9). Growing cells partly attached to the glass walls. The guanine plus cytosine content of the DNA was 40.9 mol%. Of several single compounds (H_2; alkanes including isobutane; primary and secondary alcohols; alkanoates from C_1 through to C_6; lactate; fumarate; succinate) tested, strain BuS5 used only propane and *n*-butane. During incubation with a mixture of methane, ethane, propane and butane, strain BuS5 consumed only propane and butane (Fig. 2). Because there was no pronounced exponential growth phase,

[1]Max Planck Institute for Marine Microbiology, Celsiusstraße 1, D-28359 Bremen, Germany. [2]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02536, USA. [3]GeoForschungsZentrum Potsdam, Telegrafenberg, D-14473 Potsdam, Germany. [4]Institute of Biogeochemistry and Marine Chemistry, Bundesstraße 55, University of Hamburg, 20146 Hamburg, Germany. [5]Institute for Organic Chemistry, RWTH Aachen University, D-52056 Aachen, Germany. [6]Department of Marine Sciences, University of Georgia, Athens, Georgia 30602-3636, USA. †Present address: Leibniz Institute for Natural Product Research and Infection Biology, Hans Knöll Institute, D-07745 Jena, Germany.

Table 1 | Anaerobic sulphate-reducing enrichment cultures obtained with gaseous hydrocarbons

Substrate	Cultivation temperature (°C)	Inoculum (sediment)	Culture designation*	Characteristic morphotype(s)	Dominant phylotype†
Ethane	12	Gulf of Mexico	None	Not yet evident in early state	Not identified
Propane	12	Gulf of Mexico	None	Rod-shaped cells	Not identified
Propane	28	Guaymas basin	None	Rod-shaped cells	Not identified
Propane	60	Guaymas basin	Propane60-GuB	Spindle-shaped cells	*Desulfotomaculum*
n-Butane	12	Gulf of Mexico	Butane12-GMe	Oval and curved cells	DSS cluster‡
n-Butane	28	Guaymas basin	Butane28-GuB	Oval cells, like isolated strain BuS5	DSS cluster‡
n-Butane	60	Guaymas basin	None	Various rod-shaped cells	Novel cluster§

* For convenience, a designation was used for those cultures that appear in Fig. 1 and Supplementary Fig. 2.
† Identified via 16S rRNA-targeted whole cell hybridization with specifically designed probes (Supplementary Table 1).
‡ *Desulfosarcina/Desulfococcus* cluster, a group of widespread marine SRB.
§ All bacterial clones ($n = 38$) were closely related to a clone (a1b0202) from a 16S rRNA-based survey of sediment[30] that had been freshly retrieved from Guaymas basin during the same cruise; these clones may be the representatives of a novel bacterial lineage. On the other hand, archaeal clones were also detected that showed a higher diversity and affiliated with *Thermococcus mexicalis, Archaeoglobus profundus* and relatives of *Thermoplasma, Methanosaeta* and *Methanococcus*, which are unlikely to have a direct role in butane degradation. However, growth in this culture was accompanied by signs of lysis of a proportion of the cells and possible development of 'secondary feeders', so that hybridization results must be viewed critically.

a doubling time of 4–5 days could be estimated only roughly from an early phase of the sulphide production curve.

Substrate and product quantification (Supplementary Table 2) in growth experiments with the thermophilic enrichment with propane (Propane60-GuB) and the mesophilic strain BuS5 with butane suggested a complete oxidation according to following respective equations:

$$2C_3H_8 + 5SO_4^{2-} + 4H^+ \rightarrow 6HCO_3^- + 5H_2S + 2H_2O \quad (1)$$

$$\Delta G^\circ{}_{pH7} = -102\ \text{kJ per mol propane or } -41\ \text{kJ per mol sulphate}$$

$$4C_4H_{10} + 13SO_4^{2-} + 10H^+ \rightarrow 16HCO_3^- + 13H_2S + 4H_2O \quad (2)$$

$$\Delta G^\circ{}_{pH7} = -138\ \text{kJ per mol } n\text{-butane or } -42\ \text{kJ per mol sulphate}$$

The indicated energy yields are far less than in the case of an oxidation with O_2, which would yield −2,094 kJ per mol propane and −2,727 kJ per mol *n*-butane.

Metabolites potentially indicative of the mode of activation of the short-chain hydrocarbons were searched for in the thermophilic enrichment culture with propane (Propane60-GuB) and strain BuS5. Anaerobic bacteria growing with higher alkanes are assumed to activate their substrates through homolytic C–H-bond cleavage at a subterminal carbon atom and addition to fumarate, yielding (1-methylalkyl)succinates[5,6,15]. Upon growth with propane, however,

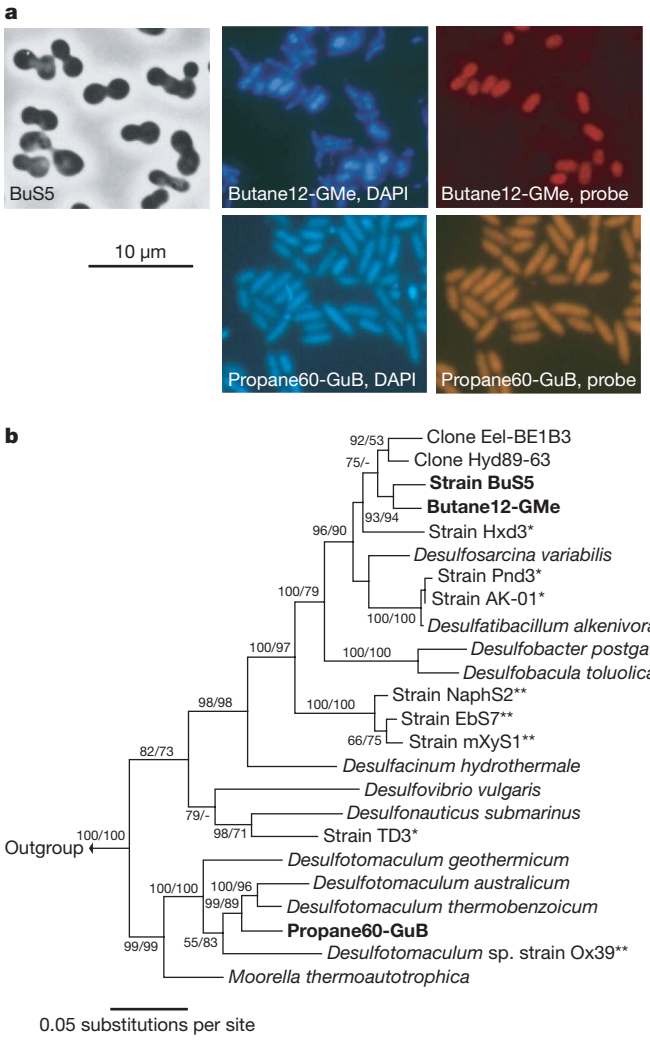


Figure 1 | Microscopy and phylogenetic analysis. **a**, Left, strain BuS5 grown with butane at 28 °C, phase contrast; middle, enrichments grown with butane at 12 °C (Butane12-GMe, top) and with propane at 60 °C (Propane60-GuB, bottom), with DAPI fluorescence; right, the same preparations as in the middle, but with fluorescence of 16S rRNA-targeted probes. **b**, Maximum-likelihood tree (16S rRNA-based) with present cultures (in bold), other bacteria and uncultured environmental phylotypes. Degraders of higher alkanes (*) and aromatic hydrocarbons (**) have been marked. Numbers indicate bootstrap values (>50%) from distance (first number) and parsimony (second number) analyses. See also Supplementary Fig. 2.

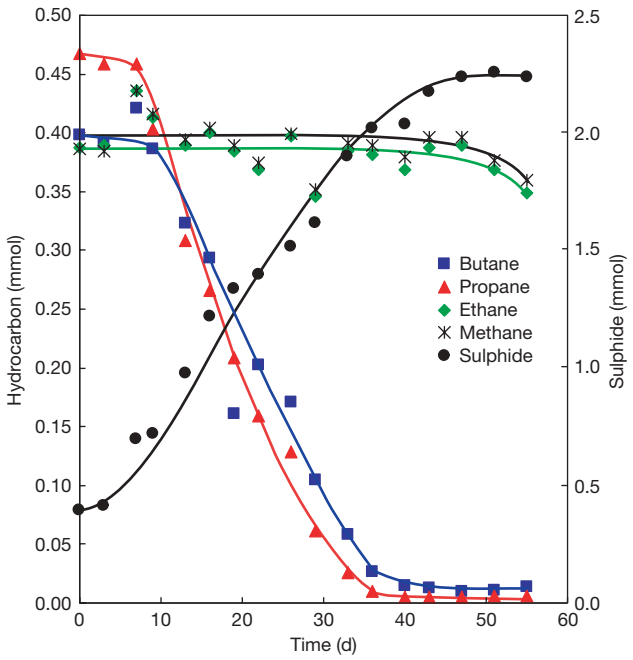


Figure 2 | Time course of the anaerobic consumption of propane and butane by a mesophilic strain. Strain BuS5 (200 ml culture, 52 ml gas space) consumes propane and *n*-butane simultaneously from a gas mixture and forms sulphide from sulphate. Methane and ethane were not degraded, but show a slight abiotic loss also observed in sterile controls. (For measurements with other cultures, see Supplementary Figs 4 and 5.)

we identified *n*-propylsuccinate in both cultures in addition to the expected isopropylsuccinate (Supplementary Fig. 6). This finding suggests two routes for propane, one starting with the common activation at the secondary carbon atom and another with the unprecedented activation at the primary carbon atom (Fig. 3). So the microorganisms in the slow, early-state enrichment with ethane (which has only primary carbon atoms) could in principle make use of the reaction principle involving fumarate. Activation at a primary carbon atom has to overcome a higher energetic barrier than activation at a secondary carbon (Supplementary Table 3), such that the former reaction may be the slower one. This assumption would offer an explanation for the observed very slow development of the more 'difficult' ethane oxidation in comparison to propane and butane oxidation. If strain BuS5 was grown with *n*-butane, only (1-methylpropyl)succinate was detected as activation product; it apparently occurred in diastereomers (data not shown), as the activation products of higher alkanes[5,15].

As the synthesis of cellular fatty acids in many alkane-degrading bacteria starts with a building block derived from the alkane

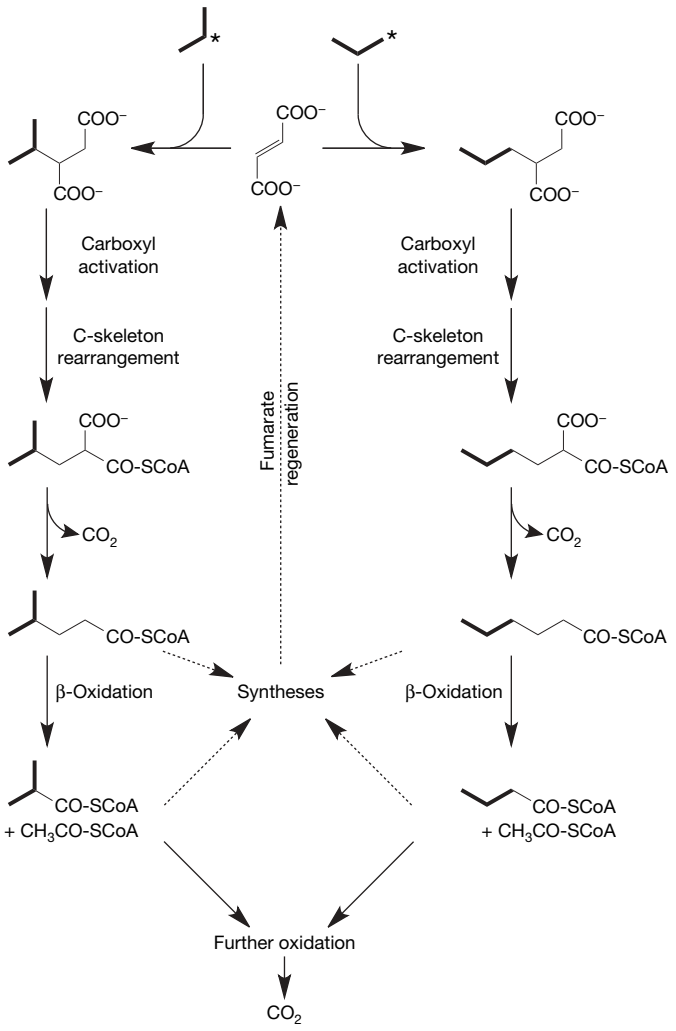


Figure 3 | Anaerobic activation reactions of propane suggested by identified metabolites. Whereas previous studies on the anaerobic degradation of alkanes suggested their activation (by homolytic C−H-bond cleavage) exclusively at a secondary carbon atom[5,6,15], the new cultures obviously activate propane at both the secondary and the primary carbon (marked with asterisks). The products from reaction with fumarate, isopropylsuccinate and *n*-propylsuccinate were identified by chemical analyses including authentic standards (Supplementary Fig. 6). Subsequent reactions are suggested according to degradation schemes of other alkanes (for details see ref. 5 and Supplementary Fig. 7) and identified fatty acids (see text and Supplementary Table 4).

metabolism[5], we also analysed the lipid fraction of strain BuS5 to further substantiate the proposed activation mechanisms (Supplementary Table 4). Strain BuS5 indeed synthesized from propane mainly *iso*-tetradecanoic acid (*i*-14:0) and *iso*-hexadecanoic acid (*i*-16:0) and from butane mainly *anteiso*-pentadecanoic acid (*ai*-15:0). The structure of these fatty acids is in accordance with the proposed further metabolism of the activation products of propane and butane, isopropylsuccinate and (1-methylpropyl) succinate, respectively, leading to 4-methylpentanoyl-CoA and 4-methylhexanoyl-CoA, respectively, as starting molecules for fatty acid biosynthesis (Supplementary Fig. 7; CoA, coenzyme A). Signatures of the additional terminal propane activation are less clear in the fatty acid patterns because the resulting C-even *n*-fatty acid (Supplementary Fig. 7) would add to the regular (substrate-independent) background of these common types of fatty acids.

Finally, we also examined whether anaerobic alkane oxidation in the obtained cultures is associated with stable carbon isotope (^{12}C versus ^{13}C) fractionation; such fractionation *in situ* is regarded as an important indicator of biodegradation. Indeed, the cultures analysed so far preferentially used the isotopically lighter hydrocarbons. The isotope ratio of propane and butane in sterile controls remained unaffected. During propane degradation by the cold-adapted enrichment culture (no designation), the mesophilic strain BuS5, and the thermophilic enrichment culture (Propane60-GuB), we observed fractionation factors (α_C) of 1.0059, 1.0052 and 1.0059, respectively. During butane degradation, the cold-adapted enrichment culture (Butane12-GMe) and the mesophilic strain BuS5 both exhibited a fractionation factor of 1.0016 (Supplementary Table 5). The fractionation factor for the anaerobic oxidation of methane in marine sediments is between 1.0088 and 1.011 (refs 16−18). Obviously, the extent of isotopic discrimination decreases with the molecular size of the alkane substrate, as also observed during their aerobic degradation[19].

Our present findings offer an explanation at the organism level of several *in situ* phenomena previously attributed to an assumed biological oxidation of propane and butane. In the marine environment, such phenomena are, for instance: the obvious disappearance of ethane, propane and butane together with methane in mud volcanoes (Gulf of Cadiz)[10]; the ^{13}C-enriched propane and butane in sediment interstitial water in comparison to propane and butane in parental gas hydrates[9]; the carbonate alkalinity around gas hydrates (Gulf of Mexico) largely attributed to oxidation of propane and butane[20]; the insufficient explanation of bulk sulphate reduction at gas seeps (Gulf of Mexico, Guaymas basin) solely by methane oxidation, and postulation of other hydrocarbons serving as electron donors[11,21]; the formation of so-called dry gas[8] (methane-enriched) caps above biodegraded oil rims due to biodegradation of short-chain (non-methane) hydrocarbons, in particular propane and butane[22]; and the suspected nourishment by short-chain hydrocarbons of microbial communities in serpentinite-hosted hydrothermal systems, such as the Lost City hydrothermal field[23]. The activities of such organisms could potentially also lead to destabilization of structure II gas hydrates[20]. On a global scale, these microbial activities may influence the emission of non-methane hydrocarbons from the oceans[24]. Our culture data also substantiate findings in terrestrial environments, such as detection of metabolites of short-chain hydrocarbons in polluted aquifers[6] and the supposed support by short-chain hydrocarbons of a deep subsurface community that included molecular signatures of *Desulfotomaculum*[25]. A growing number of studies also indicates the preferential degradation of propane and butane in subsurface gas reservoirs, a process that would render the geochemical composition-based evaluation of the gas origin and migration more difficult[7,8]. According to our cultures, phylogenetically and phenotypically diverse SRB belonging to the Deltaproteobacteria and Gram-positive bacteria are able to utilize propane and butane. In view of the limitations of cultivation techniques, an even greater diversity can be

expected to be responsible for oxidation of short-chain hydrocarbons *in situ*.

METHODS SUMMARY

Anoxic sediment samples were collected at hydrocarbon seep areas in the Gulf of Mexico at 550 m water depth and the Guaymas basin (Gulf of California) at 2,000 m water depth.

Cultures were enriched and grown at 12, 28 or 60 °C anaerobically in defined synthetic seawater medium[5] with 28 mM sulphate in stoppered bottles (usually 100 or 200 ml) or tubes (20 ml) under an anoxic atmosphere (one-third of the bottle volume) containing the hydrocarbon gas, nitrogen and 10% CO_2. For subcultivation, 10% of the culture was transferred to fresh medium of the same composition. The pure culture of strain BuS5 was obtained via dilution in anoxic agar tubes with butane in the head space.

Isolation of DNA and sequencing of amplified 16S rRNA genes (~1,500 bp) was performed according to standard protocols. Sequences were phylogenetically analysed using the ARB software[26–28]. The phylogenetic tree was inferred from maximum likelihood analysis.

Fluorescence whole-cell hybridization with specifically designed Cy3-labelled oligonucleotide probes (Supplementary Table 1) and determination of total cell counts using DAPI was carried out as described[29].

Sulphide was quantified colorimetrically in a reaction yielding methylene blue. Gaseous hydrocarbons were quantified using a gas chromatograph with a flame ionization detector. Metabolites were extracted, converted to methyl esters and analysed by gas chromatography–mass spectrometry[15]. The included standards of *n*-propylsuccinic and isopropylsuccinic acid methyl esters were synthesized using the respective alkylmercuric acetates, fumaric acid dimethyl ester and sodium borohydride. Bacterial fatty acids were analysed as methyl esters by gas chromatography–mass spectrometry. Carbon isotope ratios of propane and butane were analysed through coupling of gas chromatography, combustion and isotope ratio mass spectrometry.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 October 2006; accepted 30 August 2007.
Published online 19 September 2007.

1. Claypool, G. E. & Kvenvolden, K. A. Methane and other hydrocarbon gases in marine sediment. *Annu. Rev. Earth Planet. Sci.* **11,** 299–327 (1983).
2. Hinrichs, K.-U. *et al.* Biological formation of ethane and propane in the deep marine subsurface. *Proc. Natl Acad. Sci. USA* **103,** 14684–14689 (2006).
3. Ashraf, W., Mihdhir, A. & Murrel, J. C. Bacterial oxidation of propane. *FEMS Microbiol. Lett.* **122,** 1–6 (1994).
4. Arp, D. J. Butane metabolism by butane-grown *'Pseudomonas butanovora'*. *Microbiology* **145,** 1173–1180 (1999).
5. Widdel, F., Boetius, A. & Rabus, R. in *The Prokaryotes* 3rd edn (eds Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E.) Vol. 2, 1028–1049 (Springer, New York, 2006).
6. Gieg, L. M. & Suflita, J. M. Detection of anaerobic metabolites of saturated and aromatic hydrocarbons in petroleum-contaminated aquifers. *Environ. Sci. Technol.* **36,** 3755–3762 (2002).
7. Martini, A. M. *et al.* Microbial production and modification of gases in sedimentary basins: A geochemical case study from a Devonian shale gas play, Michigan Basin. *Bull. Am. Assoc. Petrol. Geol.* **87,** 1355–1375 (2003).
8. Wenger, L. M., Davis, C. L. & Isaksen, G. H. Multiple controls on petroleum biodegradation and impact on oil quality. *Soc. Petrol. Eng. Reserv. Eval. Eng.* **5,** 375–383 (2002).
9. Sassen, R. *et al.* Free hydrocarbon gas, gas hydrate, and authigenic minerals in chemosynthetic communities of the northern Gulf of Mexico continental slope: relation to microbial processes. *Chem. Geol.* **205,** 195–217 (2004).
10. Niemann, H. *et al.* Microbial methane turnover at mud volcanoes of the Gulf of Cadiz. *Geochim. Cosmochim. Acta* **70,** 5336–5355 (2006).
11. Joye, S. B. *et al.* The anaerobic oxidation of methane and sulfate reduction in sediments from Gulf of Mexico cold seeps. *Chem. Geol.* **205,** 219–238 (2004).
12. Orcutt, B. N. *et al.* Life at the edge of methane ice: microbial cycling of carbon and sulfur in Gulf of Mexico gas hydrates. *Chem. Geol.* **205,** 239–251 (2004).
13. Simoneit, B. R. T., Kawka, O. E. & Brault, M. Origin of gases and condensates in the Guaymas Basin hydrothermal system (Gulf of California). *Chem. Geol.* **71,** 169–182 (1988).
14. Whelan, J. K., Simoneit, B. R. T. & Tarafa, M. E. C_1-C_8 Hydrocarbons in sediments from Guaymas Basin, Gulf of California — Comparison to Peru Margin, Japan Trench and California borderlands. *Org. Geochem.* **12,** 171–194 (1988).
15. Rabus, R. *et al.* Anaerobic initial reaction of *n*-alkanes in a denitrifying bacterium: Evidence for (1-methylpentyl)succinate as initial product and for involvement of an organic radical in *n*-hexane metabolism. *J. Bacteriol.* **183,** 1707–1715 (2001).
16. Alperin, M. J., Reeburgh, W. S. & Whiticar, M. J. Carbon and hydrogen isotope fractionation resulting from anaerobic methane oxidation. *Glob. Biogeochem. Cycles* **2,** 279–288 (1988).
17. Martens, C. S., Albert, B. D. & Alperin, M. J. Stable isotope tracing of anaerobic methane oxidation in the gassy sediments of Eckernförde bay, German Baltic Sea. *Am. J. Sci.* **299,** 589–610 (1999).
18. Seifert, R., Nauhaus, K., Blumenberg, M., Krüger, M. & Michaelis, W. Methane dynamics in a microbial community of the Black Sea traced by stable carbon isotopes *in vitro*. *Org. Geochem.* **37,** 1411–1419 (2006).
19. Kinnaman, F. S., Valentine, D. L. & Tyler, S. C. Carbon and hydrogen isotope fractionation associated with the aerobic microbial oxidation of methane, ethane, propane and butane. *Geochim. Cosmochim. Acta* **71,** 272–283 (2007).
20. Formolo, M. J. *et al.* Quantifying carbon sources in the formation of authigenic carbonates at gas hydrate sites in the Gulf of Mexico. *Chem. Geol.* **205,** 253–264 (2004).
21. Kallmeyer, J. & Boetius, A. Effects of temperature and pressure on sulfate reduction and anaerobic oxidation of methane in hydrothermal sediments of Guaymas Basin. *Appl. Environ. Microbiol.* **70,** 1231–1233 (2004).
22. Head, I. M., Jones, D. M. & Larter, S. R. Biological activity in the deep subsurface and the origin of heavy oil. *Nature* **426,** 344–352 (2003).
23. Kelley, D. S. *et al.* A serpentinite-hosted ecosystem: The Lost City hydrothermal field. *Science* **307,** 1428–1434 (2005).
24. Plass-Dülmer, C., Koppmann, R., Ratte, M. & Rudolph, J. Light nonmethane hydrocarbons in seawater. *Glob. Biogeochem. Cycles* **9,** 79–100 (1995).
25. Moser, D. P. *et al. Desulfotomaculum* and *Methanobacterium* spp. dominate a 4- to 5-kilometer-deep fault. *Appl. Environ. Microbiol.* **71,** 8773–8783 (2005).
26. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32,** 1363–1371 (2004).
27. Swofford, D. L. PAUP* 4.0: phylogenetic analysis using parsimony (and other methods). (Sinauer Associates, Sunderland, Massachusetts, 1999).
28. Posada, D. & Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14,** 817–818 (1998).
29. Snaidr, J., Amann, R., Huber, I., Ludwig, W. & Schleifer, K.-H. Phylogenetic analysis and *in situ* identification of bacteria in activated sludge. *Appl. Environ. Microbiol.* **63,** 2884–2896 (1997).
30. Teske, A. *et al.* Microbial diversity of hydrothermal sediments in the Guaymas Basin: Evidence for anaerobic methanotrophic communities. *Appl. Environ. Microbiol.* **68,** 1994–2007 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are particularly indebted to the shipboard science parties and submersible operation teams of the RV *Seward Johnson II* and RV *Atlantis*, and to K. Zengler for providing sediment samples from Guaymas basin. Funding for the Gulf of Mexico cruise was provided by the US National Science Foundation (Life in Extreme Environments programme), the US Department of Energy, and the US National Oceanographic and Atmospheric Administration. We thank J. Harder, C. Karger, R. Lendt and A. Sobotta for instrumental help. S.M.S. and S.B.J. acknowledge support through a fellowship received from the Hanse Wissenschaftskolleg, Delmenhorst (Germany). This study was supported by the Max-Planck-Gesellschaft and the GEOTECHNOLOGIEN research programme of the BMBF and DFG.

Author Information The nucleotide sequences have been deposited at EMBL, GenBank and DDBJ under accession numbers EF077225 (strain BuS5), EF077226 (enrichment culture 'Butane12-GMe'), EF077227 (enrichment culture 'Propane60-GuB') and EF077228 (enrichment culture with butane at 60 °C). Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to F.W. (fwiddel@mpi-bremen.de).

METHODS

Sediment samples. Anoxic sediment samples from marine hydrocarbon seep areas were collected with research submersibles. Samples from the Gulf of Mexico were collected in the Green Canyon area at 550 m depth at sites GC232 (27° 44.4566′ N, 91° 18.981′ W) and GC234 (27° 44.7003′ N, 91° 13.3093′ W) during a cruise in July 2002 on board the RV *Seward Johnson II* (Harbor Branch Oceanographic Institution). Samples from Guaymas basin (Gulf of California) were collected during *Alvin* dive 3203 at 2,000 m depth at station 1 (27° 00.762′ N, 111° 24.656′ W) during a cruise in April/May 1998 on board the RV *Atlantis* (Woods Hole Oceanographic Institution; detailed information about the sampling site is available elsewhere[31]). Samples were stored anoxically under N_2 at 6 °C until use.

Enrichment, isolation and cultivation. Cultures were grown at 12, 28 or 60 °C in anoxic synthetic seawater medium[5,32] (CO_2/bicarbonate as buffer; sodium sulphide as reductant) with 28 mM sulphate in butyl-rubber-stoppered bottles (100–200 ml) or tubes (20 ml) under an anoxic atmosphere (one-third of bottle volume) containing the hydrocarbon gas, N_2 and 10 vol.% CO_2. For subculturing, 10% of the culture liquid was transferred. A pure culture of strain BuS5 was obtained via dilution and colony isolation using anoxic agar tubes[32] with butane in the head space.

Analyses of 16S rRNA genes. Isolation of DNA and sequencing of amplified 16S rRNA genes (~1,500 bp) was performed according to standard protocols. Sequences were aligned with those of the database of the Technical University Munich using the ARB program package[26]. Aligned sequences were imported into PAUP version 4.0b10 for phylogenetic analysis[27]. Models for minimum evolution and maximum likelihood searches were chosen using the likelihood ratio test in the MODELTEST program version 3.5 (ref. 28), with application of the GTR+I+G model. The tree was inferred from maximum likelihood analysis by heuristic searches using five random addition replicates with TBR branch swapping. Bootstrapping of minimum evolution and maximum parsimony analyses included 1,000 bootstrap replicates, each with 10 random additions.

Fluorescence hybridization. Fluorescence whole-cell hybridization with Cy3-labelled oligonucleotide probes (Biomers.net) and determination of total cell counts using DAPI (4′,6-diamidino-2-phenylindole) were carried out as previously described[29]

Probe But5-620 developed for the parental enrichment culture of the mesophilic strain BuS5 (Supplementary Fig. 3) has at least one mismatch to available 16S rRNA sequences. *Desulfococcus biacutus*, which has one mismatch, yielded only a weak signal with this probe at 50−60% formamide. *Desulfosarcina variabilis*, *Desulfofaba gelida*, and the dominant phylotype in enrichment with butane at 12 °C (Butane12-GMe), which have two mismatches, did not show signals at >10% formamide. A second probe, But5-476, substantiated the hybridization of strain BuS5 in the parental enrichment culture (not shown). Probe But5-476 is specific for strain BuS5 and its closest relative, an environmental sequence from the Gulf of Mexico (GenBank accession number AY324498; 98.3% sequence similarity). *Desulfovibrio longus*, which has two mismatches, did not show a signal with probe But5-476 at >10% formamide.

Probe K2-3-190 developed for the enrichment culture with propane at 60 °C (propane60-GuB) has at least two mismatches to available sequences. The sequences AB088961 and AF287791 (GenBank accession numbers), which have two mismatches, are from uncultivated phylotypes. *Desulfotomaculum thermobenzoicum*, which has three mismatches, did not show a signal.

Probe But12-1275 developed for the enrichment cultures with butane at 12 °C, (butane12-GMe) has at least two mismatches to available sequences. Use of the probe for this culture at 0%, 10% and 20% formamide revealed the same proportion of hybridizing cells, indicating that there were no other organisms with highly similar target regions.

The antisense probe EUB338 never yielded a signal with the presently investigated cultures, thus verifying the absence of unspecific probe binding.

Chemical analyses. Sulphide was quantified colorimetrically via reaction with N,N-dimethylphenylenediamine and subsequent oxidation yielding methylene blue[33] in a miniaturized assay[34].

Gaseous hydrocarbons were quantified on a GC14 gas chromatograph (Shimadzu) equipped with a Supel-Q PLOT (Sigma-Aldrich) fused silica capillary column (30 m × 0.53 mm) and a flame ionization detector. The carrier gas was N_2. Temperatures: column, 110 °C (isothermal); detector, 280 °C; injector, 150 °C. Standards were prepared in glass vials sealed with butyl stoppers. Volumes of 0.1 ml were withdrawn from the culture head space or the standard vials using an anoxic, gas-tight syringe (with a needle lock). Amounts of gases dissolved in the aqueous medium were taken into consideration. The partial-pressure-dependent (minor, in our experiments) dissolved proportions were calculated from determined partial pressures using molar solubility constants ($M\,atm^{-1}$): propane, 1.3×10^{-3} (30 °C) and 0.74×10^{-3} (60 °C); butane, 2.3×10^{-3} (10 °C), 1.0×10^{-3} (30 °C) and 0.52×10^{-3} (60 °C).

Metabolites were extracted with dichloromethane, converted to methyl esters and analysed by gas chromatography–mass spectrometry as described previously[16]. The included standards of *n*-propylsuccinic and isopropylsuccinic acid methyl esters were synthesized using the respective alkylmercuric acetates, fumaric acid dimethyl ester and sodium borohydride[35–37].

Cellular fatty acids were analysed as methyl esters by gas chromatography–mass spectrometry. Fatty acid methyl esters were obtained from ~20 mg cell dry mass by extraction and transesterification with trimethylchlorosilane/methanol (1/8, vol./vol.; 2 h, 70 °C), and re-extraction with *n*-hexane versus H_2O. Compounds were identified by comparison of the gas chromatographic equivalent chain lengths with those of an authentic fatty acid methyl ester mixture (37 components FAME mix; Sigma-Aldrich) as standard.

Carbon isotope ratios of propane and butane were analysed by coupling of gas chromatography, combustion, and isotope ratio mass spectrometry. Hydrocarbon gases were injected into a cooled injection system 4 (CIS4; Gerstel) operated in splitless mode. The injection system, which was packed with 10 mg PoraPak Q (80–100 mesh; Varian) was cooled to −100 °C for trapping and subsequently heated to 100 °C at 12 °C s^{-1} for desorption. The HP6890 gas chromatograph (Hewlett Packard) was equipped with a CP PoraPLOT Q capillary column (25 m × 0.32 mm, 10 μm film thickness; Varian). The temperature program was 30 °C (8 min), 5 °C min^{-1} to 150 °C (0 min), and 15 °C min^{-1} to 200 °C (5 min). The gas chromatograph was interfaced to a Delta^{Plus}XL mass spectrometer (Finnigan) equipped with a CuO/Ni/Pt combustion furnace operated at 940 °C. Fractionation factors (α_C) were calculated as described previously[18].

The DNA base ratio was determined at the DSMZ (Braunschweig, Germany) using a standard chromatographic technique.

31. Weber, A. & Jørgensen, B. B. Bacterial sulfate reduction in hydrothermal sediments of the Guaymas Basin, Gulf of California, Mexico. *Deep-sea Res. I* **49**, 827–841 (2002).
32. Widdel, F. & Bak, F. in *The Prokaryotes* 2nd edn (eds Balows, A., Trüper, H. G., Dworkin, M., Harder, W. & Schleifer, K.-H.) Vol. 4, 3352–3378 (Springer, New York, 1992).
33. Cline, J. D. Spectrophotometric determination of hydrogen sulphide in natural waters. *Limnol. Oceanogr.* **14**, 454–458 (1969).
34. Aeckersberg, F., Bak, F. & Widdel, F. Anaerobic oxidation of saturated hydrocarbons to CO_2 by a new type of sulfate-reducing bacterium. *Arch. Microbiol.* **156**, 5–14 (1991).
35. Giese, B. & Meister, J. Die Addition von Kohlenwasserstoffen an Olefine. Eine neue synthetische Methode. *Chem. Ber.* **110**, 2588–2600 (1977).
36. Giese, B. & Kretzschmar, G. Radikalische Addition an cyclische Derivate der Maleinsäure. *Chem. Ber.* **115**, 2012–2014 (1982).
37. Giese, B. & Kretzschmar, G. Radikalkettenreaktionen mit Maleinsäureanhydriden. Zur kontrathermodynamischen Stereoselektrivität. *Chem. Ber.* **117**, 3175–3182 (1984).

LETTERS

Neanderthals in central Asia and Siberia

Johannes Krause[1], Ludovic Orlando[2], David Serre[3], Bence Viola[4], Kay Prüfer[1], Michael P. Richards[1], Jean-Jacques Hublin[1], Catherine Hänni[2], Anatoly P. Derevianko[5] & Svante Pääbo[1]

Morphological traits typical of Neanderthals began to appear in European hominids at least 400,000 years ago[1] and about 150,000 years ago[2] in western Asia. After their initial appearance, such traits increased in frequency and the extent to which they are expressed until they disappeared shortly after 30,000 years ago. However, because most fossil hominid remains are fragmentary, it can be difficult or impossible to determine unambiguously whether a fossil is of Neanderthal origin. This limits the ability to determine when and where Neanderthals lived. To determine how far to the east Neanderthals ranged, we determined mitochondrial DNA (mtDNA) sequences from hominid remains found in Uzbekistan and in the Altai region of southern Siberia. Here we show that the DNA sequences from these fossils fall within the European Neanderthal mtDNA variation. Thus, the geographic range of Neanderthals is likely to have extended at least 2,000 km further to the east than commonly assumed.

The partial skeleton of an 8–10-year-old child discovered in the late 1930s in Teshik-Tash Cave, Uzbekistan, is generally accepted to represent the easternmost extent of the Neanderthal range[3]. However, its Neanderthal affinities have been disputed[4,5]. Further to the east in the Altai region of Siberia, human remains have been found in association with Mousterian lithic technology, which is usually associated with Neanderthals in Europe but is also found in association with modern humans in the Near East and northern Africa[6]. For example, teeth found at Okladnikov Cave in the Altai Mountains, which are dated between 37,750 ± 750 (1σ) and 43,700 +1,100/−1,300 years BP[7] (see also Supplementary Table 1), have been suggested to stem from Neanderthals[8]. However, others have suggested that they come from modern humans with some Asian *Homo erectus* traits[9]. Okladnikov Cave has also yielded four postcranial bones: a middle phalanx and a distal humerus fragment of adults and the distal halves of the humerus and femur from what is likely to be a single subadult individual[10]. Although the fragmentary adult humerus cannot be assigned to either Neanderthals or modern humans, the subadult remains and the adult phalanx have been suggested not to be of modern human origin[10].

To determine whether the Teshik Tash and Okladnikov individuals are genetically affiliated with European Neanderthals, we attempted to retrieve mtDNA from the left femur of Teshik Tash and the three fragmentary long bones from Okladnikov. So far, mtDNA sequences have been determined from 13 Neanderthals in Europe[11–20]. Comparison of these DNA sequences with those of mtDNAs from contemporary humans shows that the Neanderthal mtDNA gene pool was distinct from that of modern humans[16,17,21].

We extracted DNA from samples (about 200 mg) from each of the four bones and amplified DNA with the use of three different primer pairs. Each product was cloned and multiple clones were sequenced. From the Teshik Tash specimen, 88 of 90 clones from a 63-base-pair (bp) amplification product generated by primers that amplify modern human as well as Neanderthal mtDNA[22] had sequences identical to those of modern human mtDNAs, whereas two were similar to previously determined Neanderthal sequences with one additional substitution at position 16,242 (C→T) when compared with the revised Cambridge reference sequence (RCRS)[23]. A 119-bp product[22] revealed 156 clones identical to those from modern humans and no clones with similarity to Neanderthal mtDNAs, whereas a Neanderthal-specific primer pair[16] retrieved Neanderthal-like mtDNA sequences (including the substitution at position 16,242). For the subadult humerus from Okladnikov, sequences from 2 out of 104 clones from the 63-bp product were found to be identical to previously amplified Neanderthal sequences, whereas none of 103 clones from the 119-bp product was Neanderthal-like. For the subadult femur and the adult humerus, neither the shorter nor the longer products yielded any Neanderthal-like sequences. By contrast, the Neanderthal-specific primer pair retrieved products from the two subadult remains but not from the adult humerus (Supplementary Table 3).

These results show that both the subadult individual from Okladnikov Cave and the Teshik Tash individual carried mtDNA of the Neanderthal type, whereas there is no indication that the adult individual from Okladnikov did so. The high ratio of modern human DNA to Neanderthal DNA for the subadult Okladnikov and the Teshik Tash specimens are in agreement with previous observations that modern human mtDNA occurs in most fossil bones[16,24], where it often outnumbers endogenous mtDNA[22]. Next, we directly dated the adult and subadult humerus from Okladnikov. Although the adult bone yielded an uncalibrated ^{14}C date of 24,260 ± 180 years BP, the subadult bone yielded uncalibrated dates ranging from 29,990 ± 500 years BP to 37,800 ± 450 years BP (see discussion in Supplementary Information, and Supplementary Tables 1 and 2), indicating that the latter bone is old enough to be of Neanderthal origin.

We then designed primers that amplify Neanderthal mtDNA preferentially, to determine larger parts of the hypervariable region I (HVRI) from the Teshik Tash and Okladnikov specimens found to carry Neanderthal-like mtDNA. We reconstructed 190 bp of the HVRI from the Teshik Tash specimen (corresponding to positions 16,130–16,319 of the RCRS) by eight overlapping fragments in which each fragment was independently amplified at least twice, and all products were sequenced from multiple clones (Supplementary Table 6). The resulting sequence shows 22 substitutions relative to the RCRS (Supplementary Table 4). Of these, 13 have been found in all previously studied Neanderthals, six are known Neanderthal polymorphisms, and three have not previously been observed among Neanderthals (positions 16,242, 16,274 and 16,319). The remaining

[1]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. [2]Paléogénétique et Évolution Moléculaire, Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, Institut Fédératif Biosciences Gerland Lyon Sud, Université Lyon 1, CNRS, INRA, École Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon, Cedex 07, France. [3]McGill University and Genome Quebec Innovation Center, Montreal, Quebec H3A 1A4, Canada. [4]Department of Anthropology, Faculty of Life Sciences, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria. [5]Paleolithic Department, Institute of Archaeology and Ethnography, Russian Academy of Sciences, Siberian Branch. Lavrentieva Prospekt, 17 Novosibirsk, 630090 Russia.

HVRI segments from the Teshik Tash specimen could not be determined because it was not possible to design Neanderthal-specific primers spanning the entire HVRI given the short length of amplifications possible from this extract.

From the Okladnikov subadult humerus, the complete HVRI sequence between positions 16,035 and 16,378 was determined by using 16 overlapping fragments (Supplementary Table 7). It carries 22 substitutions relative to the RCRS (Supplementary Table 4), 16 of which are shared with all previous sequenced Neanderthals, four are known polymorphisms among Neanderthals, and two (positions 16,172 and 16,263) have not been seen before.

To ensure the reproducibility of the DNA sequences, bone samples of the Teshik Tash child and the Okladnikov subadult were sent to a DNA laboratory in Lyon that specializes in the analysis of ancient DNA and that did not have access to the results obtained in Leipzig. There, extracts were prepared and 48 bp of the HVR1 spanning positions 16,231–16,277 were amplified in two overlapping fragments. This HVR1 segment contains sequence positions that are common to all Neanderthals analysed so far, for example a Neanderthal-specific insertion at position 16,263b, as well as positions unique both to the Teshik Tash individual (16,242 and 16,274) and the Okladnikov individual (16,263), respectively. The consensus sequences of 67 clones for Teshik Tash and 51 clones for Okladnikov sequenced in Lyon (Supplementary Table 8) were identical to the corresponding sequences determined in Leipzig.

The HVR1 sequences from Okladnikov together with seven other Neanderthal sequences that span at least positions 16,078 to 16,378 (see Methods) were used to estimate the mean pairwise mtDNA sequence difference between Neanderthals across their entire currently known geographical range. Whereas for Neanderthals this difference is found to be 5.5 substitutions, it is 8.1 substitutions for 1,467 contemporary Africans, 6.3 substitutions for 3,217 Asians and 4.0 for 2,667 contemporary Europeans (Supplementary Table 5). Thus, mtDNA diversity in Neanderthals seems to have been within the range of present-day Europeans and Asians but significantly lower ($P = 0.0255$) than the diversity of present-day Africans (Supplementary Fig. 2).

The similarity of the Okladnikov and Teshik Tash mtDNA sequences to mtDNA determined from Neanderthals in Europe and the Caucasus, in conjunction with the absence of Neanderthal-like mtDNA from the more than 10,000 modern humans studied so far as well as from early modern humans[16,25], indicates that the Teshik Tash and Okladnikov individuals belonged to a population related to European and western Asian Neanderthals. This agrees with morphological evidence that the Teshik Tash hominid is of Neanderthal origin[3] and the suggestion that the subadult Okladnikov individual is related to Neanderthals on the basis of the morphology of teeth found in association with the bones[8]. The geographical range of Neanderthals therefore seems to have extended at least 2,000 km farther east into southern Siberia than has generally been assumed (Fig. 1).

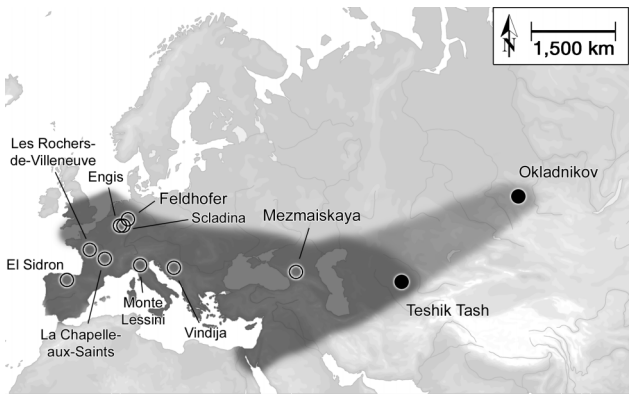


Figure 1 | Geographical range of Neanderthals. The previously known Neanderthal range based on the morphology of fossils[30] is indicated in dark grey; the Neanderthal range based on mtDNA is indicated in light grey. Sites where mtDNA sequences of the Neanderthal type were detected previously are shown as open circles, and the two sites presented in this study are indicated by black dots.

When the relationship of the Teshik Tash and Okladnikov mtDNA sequences to other Neanderthal mtDNA sequences is estimated (Supplementary Fig. 3), the Teshik Tash mtDNA sequence seems to be more closely related to the mtDNA sequence from Scladina in western Europe than to the sequence from Okladnikov. Further Neanderthal mtDNA sequences from across their range and from different time horizons will obviously be necessary to permit insights into how Neanderthals colonized both western and central parts of the Old World. However, the fact that no deep mtDNA divergence is seen between the central Asian Neanderthals and European and Caucasian Neanderthals shows that they were not separated for a long time. This supports the view that central Asia was colonized relatively recently by Neanderthals[26]. In fact, it has been suggested that Neanderthals did not colonize most of the Russian plains before an exceptionally warm episode 125,000 years ago[27], during which the Caspian Sea was drastically reduced in size. This may have facilitated the expansion of Neanderthals into central Asia and Southern Siberia[26]. Intriguingly, their presence in southern Siberia raises the possibility that they may have been present even farther to the east, in Mongolia and China. Further work will be necessary to address this possibility.

METHODS SUMMARY

DNA extraction was performed in a laboratory dedicated exclusively to work on ancient DNA. About 200 mg of bone was powdered and extracted as described[28]. To minimize the loss of material, a two-step multiplex polymerase chain reaction (PCR)[29] in a total volume of 20 μl containing up to nine primer pairs was performed. In total, 60 cycles of PCR were completed, 27 in the first step and 33 in the second. All reaction conditions were as described except for the annealing temperature, which was 55 °C for all primer pairs used. Amplification products of the correct size were cloned with the TOPO TA cloning kit (Invitrogen), and 3–17 clones were sequenced with an ABI3730 capillary sequencer (Applied Biosystems) for each product. In total, nine primer pairs in overlapping fragments with a length of 61–85 bp were used to reconstruct the 192 bp of the Teshik Tash sequence (Supplementary Table 6). For the 345 bp of the Okladnikov subadult humerus HVR1 sequence, 17 primer pairs were used, ranging from 61 to 109 bp (Supplementary Table 7). Every position was amplified independently at least twice, and for a third time where a difference between all clones from one amplification and all clones from another amplification was observed. Amplification and sequencing procedures in the Lyon laboratory are given in Supplementary information.

Several previously published Neanderthal HVR1 sequences that span at least positions 16,078–16,378 of the RCRS and Okladnikov (subadult humerus) were used to calculate the mean pairwise difference as a measure of mtDNA diversity within Neanderthals. Sequence alignments of the 2,667 European, 3,217 Asian and 1,467 African modern human HVR1 sequences were retrieved from the HvrBase++ site (www.HVRbase.org) and restricted to positions 16,078–16,378. The mean pairwise difference for the eight Neanderthal and three modern human HVR1 sequence alignments was calculated with the software MEGA3.1. From each of the three modern human populations, eight sequences were randomly extracted without replacement and the mean of their pairwise differences was calculated 10,000 times (Supplementary Table 5). The mean pairwise difference among the Neanderthal sequences falls in the lower 5% ($P = 0.0255$) of the values observed for Africans (Supplementary Fig. 2).

Received 15 March; accepted 23 August 2007.
Published online 30 September 2007.

1. Stringer, C. B. & Hublin, J. New age estimates for the Swanscombe hominid, and their significance for human evolution. *J. Hum. Evol.* **37**, 873–877 (1999).
2. Grun, R. & Stringer, C. Tabun revisited: revised ESR chronology and new ESR and U-series analyses of dental material from Tabun C1. *J. Hum. Evol.* **39**, 601–612 (2000).
3. Debetz, G. The anthropological features of the human skeleton from the cave of Teshik-Tash [in Russian]. *Trudy Uzbekist. Fil. Akad. Nauk.* **1**, 46–49 (1940).
4. Weidenreich, F. The Paleolithic child from the Teshik-tash Cave in Southern Uzbekistan (Central Asia). *Am. J. Phys. Anthropol.* **3**, 151–162 (1945).
5. Glantz, M. M. & Ritzman, T. B. A reanalysis of the Neandertal status of the Teshik-Tash child. *Am. J. Phys. Anthropol.* 38 (suppl.), 100–101 (2004).

6. Finlayson, C. & Carrion, J. S. Rapid ecological turnover and its impact on Neanderthal and other human populations. *Trends Ecol. Evol.* **22,** 213–222 (2007).
7. Derevianko, A. P. *To the Problem of Neandertal Habitation of Central Asia and Siberia* (Institute of Archaeology and Ethnography Press, Novosibirsk, 2007).
8. Turner, C. G. in *Chronostratigraphy of the Paleolithic in North, Central, East Asia and America* (ed. Derevianko, A. P.) 239–243 (USSR Academy of Sciences, Novosibirsk, 1990).
9. Shpakova, E. G. & Derevianko, A. P. The interpretation of Odontological Features of Pleistocene Human Remains from the Altai. *Archaeol. Ethnol. Anthropol. Eurasia N* **1,** 125–138 (2000).
10. Viola, T. B. *et al.* in *Terra Nostra 2006/2 150 Years of Neanderthal Discoveries* 139 (GeoUnion Alfred-Wegner-Stiftung, Berlin, 2006).
11. Lalueza-Fox, C. *et al.* Mitochondrial DNA of an Iberian Neandertal suggests a population affinity with other European Neandertals. *Curr. Biol.* **16,** R629–R630 (2006).
12. Caramelli, D. *et al.* A highly divergent mtDNA sequence in a Neandertal individual from Italy. *Curr. Biol.* **16,** R630–R632 (2006).
13. Orlando, L. *et al.* Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence. *Curr. Biol.* **16,** R400–R402 (2006).
14. Lalueza-Fox, C. *et al.* Neandertal evolutionary genetics: mitochondrial DNA data from the Iberian peninsula. *Mol. Biol. Evol.* **22,** 1077–1081 (2005).
15. Beauval, C. *et al.* A late Neandertal femur from Les Rochers-de-Villeneuve, France. *Proc. Natl Acad. Sci. USA* **102,** 7085–7090 (2005).
16. Serre, D. *et al.* No evidence of neandertal mtDNA contribution to early modern humans. *PLoS Biol.* **2,** 313–317 (2004).
17. Krings, M. *et al.* A view of Neandertal genetic diversity. *Nature Genet.* **26,** 144–146 (2000).
18. Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90,** 19–30 (1997).
19. Ovchinnikov, I. V. *et al.* Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404,** 490–493 (2000).
20. Schmitz, R. W. *et al.* The Neandertal type site revisited: Interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc. Natl Acad. Sci. USA* **99,** 13342–13347 (2002).
21. Currat, M. & Excoffier, L. Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biol.* **2,** e421 (2004).
22. Green, R. E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444,** 330–336 (2006).
23. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genet.* **23,** 147 (1999).
24. Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Pääbo, S. Ancient DNA. *Nature Rev. Genet.* **2,** 353–359 (2001).
25. Caramelli, D. *et al.* Evidence for a genetic discontinuity between Neandertals and 24,000-year-old anatomically modern Europeans. *Proc. Natl Acad. Sci. USA* **100,** 6593–6597 (2003).
26. Hublin, J.-J. in *Neandertals and Modern Humans in Western Asia* (eds Akazawa, T., Aoki, K. & Bar-Yosef, O.) 295–310 (Plenum, New York, 1998).
27. Hoffecker, J. F. *Desolate Landscapes: Ice-Age settlement in Eastern Europe* (Rutgers Univ. Press, New Brunswick, NJ, 2002).
28. Rohland, N. & Hofreiter, M. Comparison and optimization of ancient DNA extraction. *Biotechniques* **42,** 343–352 (2007).
29. Krause, J. *et al.* Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* **439,** 724–727 (2006).
30. Stringer, C. & Andrews, P. *The Complete World of Human Evolution* 156 (Thames & Hudson, London, 2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank V. M. Kharitonov for the Teshik Tash fossil samples; A. Briggs, E. Green, M. Hofreiter and C. Lalueza-Fox and members of the Max Planck Institute evolutionary genetics department for helpful comments; to B. Höbner, B. Höffner and A. Weihmann for sequencing; S. Keats for establishing contacts during the initial phase of the project; V. Wiebe for interpreting during our visits to Russia and Uzbekistan; and K. Finstermeier for help with figures. We acknowledge the Russian Academy of Sciences and its Siberian Branch for logistic support, and the Max Planck Society for funding.

Author Contributions A.D. provided Neanderthal samples and palaeontological information; S.P. and D.S. collected the samples; B.V. and J.J.H. provided palaeontological and archaeological information; M.P.R. performed dating; J.K., L.O. and D.S. extracted ancient DNA; J.K. and L.O. amplified and sequenced DNA; J.K. performed the phylogenetic analyses and the statistical analysis in cooperation with K.P.; C.H. coordinated the work in Lyon; S.P. initiated, planned and coordinated the study; J.K. and S.P. wrote the paper.

Author Information The Teshik Tash and Okladnikov Neanderthal sequences are deposited in GenBank under accession numbers EU078679 and EU078680, respectively. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.P. (paabo@eva.mpg.de).

LETTERS

Early human use of marine resources and pigment in South Africa during the Middle Pleistocene

Curtis W. Marean[1], Miryam Bar-Matthews[3], Jocelyn Bernatchez[2], Erich Fisher[4], Paul Goldberg[5], Andy I. R. Herries[6], Zenobia Jacobs[7], Antonieta Jerardino[8], Panagiotis Karkanas[9], Tom Minichillo[10], Peter J. Nilssen[11], Erin Thompson[1], Ian Watts[12] & Hope M. Williams[2]

Genetic and anatomical evidence suggests that *Homo sapiens* arose in Africa between 200 and 100 thousand years (kyr) ago[1,2], and recent evidence indicates symbolic behaviour may have appeared ~135–75 kyr ago[3,4]. From 195–130 kyr ago, the world was in a fluctuating but predominantly glacial stage (marine isotope stage MIS6)[5]; much of Africa was cooler and drier, and dated archaeological sites are rare[6,7]. Here we show that by ~164 kyr ago (±12 kyr) at Pinnacle Point (on the south coast of South Africa) humans expanded their diet to include marine resources, perhaps as a response to these harsh environmental conditions. The earliest previous evidence for human use of marine resources and coastal habitats was dated to ~125 kyr ago[8,9]. Coincident with this diet and habitat expansion is an early use and modification of pigment, probably for symbolic behaviour, as well as the production of bladelet stone tool technology, previously dated to post-70 kyr ago[10–12]. Shellfish may have been crucial to the survival of these early humans as they expanded their home ranges to include coastlines and followed the shifting position of the coast when sea level fluctuated over the length of MIS6.

The Middle Stone Age (MSA) was the technological phase (appearing as early as 280 kyr ago[13]) of the origins of modern humans. South Africa has provided a rich and important MSA archaeological record that mostly post-dates 120 kyr ago, but many of the excavated sites occur at elevations at which the MIS5e high sea stand (~+6 m above mean sea level at 123 kyr ago) would have washed out earlier deposits[14]. Co-occurring with an increase in the frequency of known coastal sites younger than 120 kyr is a burgeoning of material cultural complexity. In addition to the later appearance of bone tools[15] and beads[3], there is extensive evidence for systematic use of pigment by 120 kyr ago in South Africa[16]. Pigments near this age, or older, are patchily distributed outside South Africa; for example, in Israel some are dated to 92 kyr ago[17], at Twin Rivers in Zambia they are dated to between 141 and >400 kyr ago[18], and 'red ochres' from the Kapthurin Formation in Kenya are dated to >285 kyr ago[7].

Site PP13B (S34° 12′ 44″ E22° 05′ 37″) is a sea cave overlooking the Indian Ocean in the quartzitic coastal cliffs at Pinnacle Point near Mossel Bay (South Africa). It escaped the MIS5e high sea stand by virtue of its elevation (+15 m above mean sea level, see Supplementary Information and Video). On the north and south walls of the cave is a variably cemented MSA deposit (called the LC-MSA), whereas the floor is covered by a mostly uncemented MSA deposit (Fig. 1)[19]. We excavated three areas in the cave, but our focus here is on the LC-MSA deposits, which are the oldest sediments with significant anthropogenic input.

We recognize the following stratigraphic sequence in the LC-MSA deposits, from bottom to top. The LC-MSA Lower has a weighted mean optically stimulated luminescence (OSL) age of 164 ± 12 kyr (Table 1 and Supplementary Information, for all ages that we present). It is a sandy deposit, the least cemented of all, with multiple lenses of burnt carbonaceous material; micromorphology and frequency-dependent magnetic susceptibility analysis show that some of these are *in situ* combustion features. Lithics and faunal remains are abundant throughout, but are concentrated in burnt lenses. The plotted finds are inclined mostly between 0 and 15°, with few dramatic changes that would indicate intrusions or disturbances (Supplementary Information). The LC-MSA Middle (weighted-mean OSL age of 132 ± 12 kyr) is mostly ash, has multiple lenses of dark organic material that micromorphology shows are *in situ* hearths, and has a plotted find density less than the LC-MSA Lower. The LC-MSA Upper is a heavily cemented zone with three sub-units: (1) a lower hard sandy and silty layer containing reworked ashes that directly contacts and transitions into the richer anthropogenic deposits of the LC-MSA Middle; (2) a layer of shellfish stratified within an aeolian dune with a weighted mean OSL age of 120 ± 7 kyr; and (3) a dune that sealed the cave, and which has a weighted mean OSL age of 90 ± 6 kyr. Capping the deposit, the LC-MSA Flowstone is a 5-cm-thick laminated flowstone with a rough and wavy microscopically sharp boundary with the underlying LC-MSA Upper. Six uranium series (U-series) ages on separate laminae range from 39 to 92 kyr (Table 1), providing high resolution minimum ages for everything below. Several forms of evidence suggest partial closure of the cave from 39 to 92 kyr ago (Supplementary Information).

The LC-MSA Lower falls within MIS6, during which sea level was lower than today. We have developed a three-dimensional Geographical Information Systems (GIS) model (Supplementary Video and Information) that joins offshore bathymetry to a relative sea level curve[20] and models the coastline at 1.5-kyr increments for the last 400 kyr. Studies of modern and recent archaeological shellfish transport show that foragers rarely transport shell over more than 5–10 km[9,21]. Our model shows that the coastline was within that distance during MIS6 only at 167 kyr ago—a result concordant with the OSL ages.

The flaked-stone-artefact assemblage ($n = 1836$) is quartzite dominated (78%) and includes Levallois technology, often considered characteristic of the MSA[22], as well as bladelet technology, which is more typical of much later periods[10,11](Fig. 2). The blades form a continuous size distribution of widths from large blades to bladelets. Bladelets here conform to the formal definition of blades less than

[1]Institute of Human Origins, [2]School of Human Evolution and Social Change, PO Box 872402, Arizona State University, Tempe, Arizona 85287-2402, USA. [3]Geological Survey of Israel, 30 Malchei Israel Street, Jerusalem 95501, Israel. [4]Department of Anthropology, University of Florida, Gainesville, Florida 32611, USA. [5]Department of Archaeology, Boston University, 675 Commonwealth Avenue, Boston, Massachusetts 02215, USA. [6]Human Origins Group, School of Medical Sciences, The University of New South Wales, Sydney NSW 2052, Australia. [7]School of Earth and Environmental Sciences, University of Wollongong, Wollongong, 2522, Australia. [8]Department of Archaeology, University of Cape Town, Rondebosch 7701, South Africa. [9]Ephoreia of Palaeoanthropology-Speleology, Ministry of Culture, Ardittou 34b, Athens 11636, Greece. [10]Department of Anthropology, University of Washington, Box 353100, Seattle, Washington 98195-3100, USA. [11]Archaeology Division, Iziko-South African Museum, PO Box 61, Cape Town 8000, South Africa. [12]58 Eastdown House, Downs Estate, Amhurst Road, London E8 2AT, United Kingdom.

10 mm in width. Thirty-five plotted finds meet this definition and an additional 29 bladelets or bladelet fragments were recovered from the 3-mm screens (Supplementary Information). Bladelet technology is a significant component of the lithic assemblage: for example, in the LC-MSA Lower the total number of true bladelets ($n = 64$) exceeds the total number of Levallois products ($n = 47$).

There are 57 pigment pieces (93.4 g total) and most are from the LC-MSA Lower. Forty-six are iron-rich fine-grained sedimentary materials, and most have a pinkish-brown or reddish-brown surface colour. Streak colour (Natural Colour System) shows the majority ($n = 31$) as intermediate reddish-brown, followed by saturated reddish-brown ($n = 10$), and saturated very red ($n = 7$, high chroma values and ≥75% redness). All can be classified as 'red ochre'. Ten pieces were definitely used (eight ground and two scraped) and two pieces were probably used (both ground). Most ground pieces are moderately to intensively ground on one principal surface (Fig. 2). Saturated very-red values are disproportionately represented among used pieces, suggesting preferential use of the reddest, most chromatic ochre (Supplementary Information).

Fifteen categories of marine invertebrates are so far documented in the PP13B deposits (Table 2): four categories to species level (*Perna perna, Choromytilus meridionalis, Scutellastra argenvillei* and *Turbo sarmaticus*), five to genus level (*Donax* spp., *Helcion* sp., *Oxystele* spp., *Nodilittorina* spp., *Burnupena* spp.), three to Family level (Mytilidae, Patellidae and Turritellidae), one to Subphylum (shore barnacle, Crustacea), and two to only molluscs (whelks and chitons), to which, respectively, many families and an entire class belong. *P. perna* (brown mussel) is dominant, followed by *T. sarmaticus* (giant periwinkle), limpets (*S. argenvillei* and Patellidae) and small numbers of whelks. Brown mussels are the overwhelmingly dominant species in the LC-MSA Upper. On the basis of current habitats[23], the vast majority of shellfish were collected from exposed to moderate rocky shores and from tidal pools, easily achieved during daily low tides and/or monthly spring low tides. The whale barnacle fragment is tentatively identified as *Coronula diadema*[24], and suggests scavenging of beached whale blubber and skin with attached barnacles[25].

For millions of years, hominin diet was restricted to terrestrial plants and animals. The expansion to shellfish is one of the last

Table 1 | Radiometric ages from the LC-MSA at PP13B

U-series ages for the LC-MSA flowstone

Sample number	[U] (p.p.m.)	$^{234}U/^{238}U$	$^{230}Th/^{232}Th$	Uncorrected age (kyr)	Corrected age (kyr)	±2 σ (kyr)
32204	1.2	2.58712	186	39.1	39.1	0.4
32203	0.06	2.47122	66	45.9	45.5	0.4
32202	0.97	2.39069	23.2	47.4	46.1	1.3
32200	0.495	1.84941	46.3	75	74.03	0.8
32201	0.36	2.39069	46	81.5	80.5	1.4
32205	0.79	2.05105	42.2	92.7	91.6	1

Optically stimulated luminescence ages on the LC-MSA sediments

Sample number	Stratigraphic position	Age (kyr)	±2 σ (kyr)
46447	LC-MSA Upper (upper dune)	88.3	10.0
111400	LC-MSA Upper (upper dune)	89.3	8.2
46467	LC-MSA Upper (lower dune)	119.0	8.8
20720	LC-MSA Upper (lower dune)	121.3	8.4
111401	LC-MSA Upper (lower dune)	116.7	10.6
111402	LC-MSA Middle	135.2	12.8
111403	LC-MSA Lower	161.9	15.2
20721	LC-MSA Lower	167.7	18.2
111406	LC-MSA South Profile (lower)	161.7	17.0

Ages are ordered stratigraphically from top to bottom.

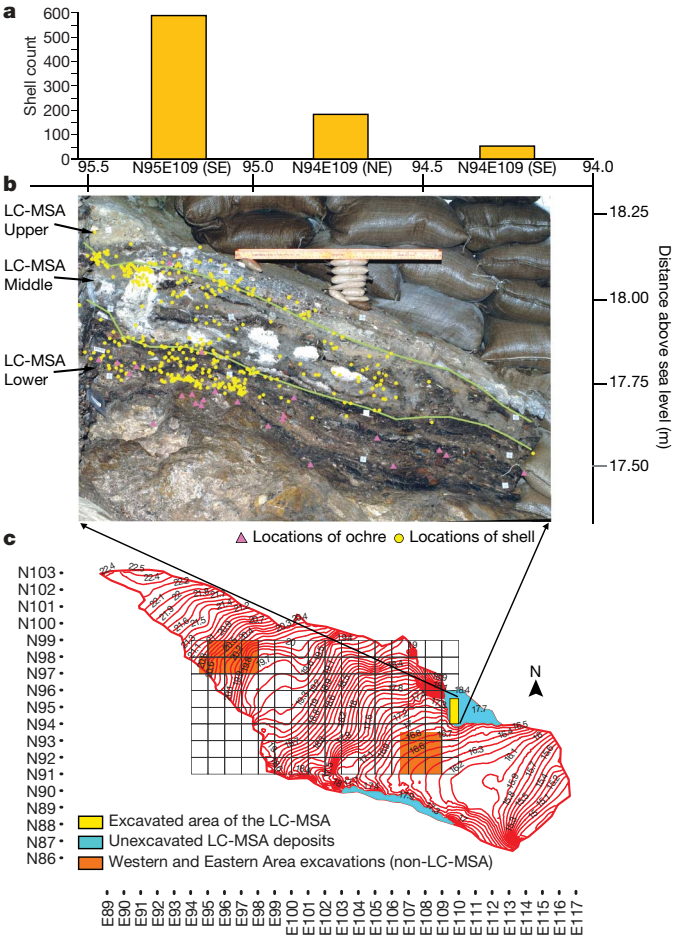


Figure 1 | Contour map of PP13B, photograph of the LC-MSA eastern section, and the frequency of shellfish and ochre. **a**, The number of shellfish per 50 cm × 50 cm excavated quadrant (NE, north-east quadrant, SE, south-east quadrant; N95E109, 95 metres north and 109 metres east of the zero point; for explanation of MAP grid space see Methods) for all layers north to south. **b**, A geo-referenced section photograph of the eastern section with plotted ochre and shellfish on the photograph (see Supplementary Information for an un-geo-referenced version), and the main divisions of the layers indicated. **c**, A contour map of 13B showing the location of the LC-MSA deposits and the positions of the excavations within the MAP grid.

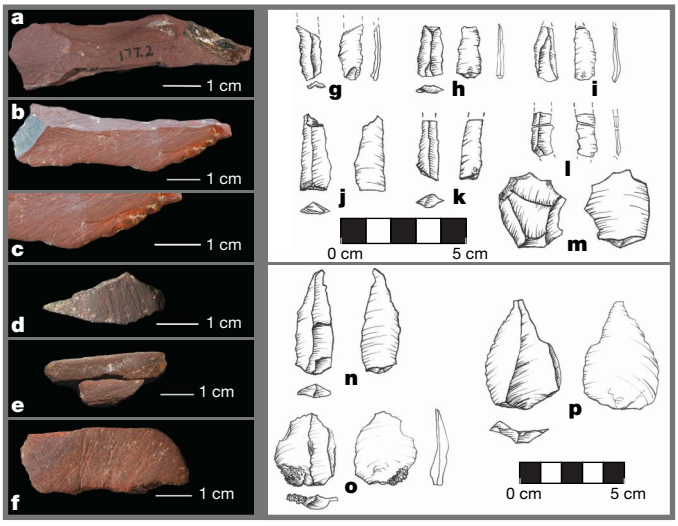


Figure 2 | Ochre and lithics from the LC-MSA Lower. **a**–**c**, Specimen 177.2 (moderately ground shale, moderately haematized, 17.1 g, NCS 3355 Y70R) three views: **a**, flaked surface. **b**, ground surface. **c**, close-up of ground surface. **d**, Specimen 79092 (ground fragment, haematite, 6 g, NCS 6727 Y75R). **e**, Specimens 81614 and 81681 (two conjoining ground fragments, siltstone, 2 g, NCS 3162 Y60R). **f**, Specimen 81770 (Intensively ground fragment, coarse siltstone, moderately haematized, 10.1g., NCS 3257 Y80R). **g**–**i**, Quartzite bladelets. **j**, Quartzite small blade. **k**, Quartzite bladelet. **l**, Quartzite bladelet fragments, refit. **m**, Quartzite core rejuvenation flake. **n**, Quartzite Levallois blade. **o**, Silcrete Levallois flake. **p**, Quartzite Levallois point.

Table 2 | The MNI and weight of shellfish in the LC-MSA layers

	LC-MSA Lower		LC-MSA Middle		LC-MSA Upper	
Species collected as food	MNI	Weight (g)	MNI	Weight (g)	MNI	Weight (g)
Mollusca						
Perna perna	14	118.5	39	310.0	20	157.8
Choromytilus meridionalis	1	2.9	1	3.1	0	0.0
Mytilidae	1	2.0	1	7.6	0	0.0
Donax spp.	1	1.1	0	0.0	0	0.0
Scutellastra argenvillei	0	0.0	1	5.1	0	0.0
Patellidae	1	5.8	5	27.0	1	0.1
Oxystele spp.	1	0.1	0	0.0	0	0.0
Turbo sarmaticus	1	9.4	4	30.7	0	0.0
Burnupena spp.	1	0.2	1	0.0	1	0.1
Whelk	1	0.8	1	1.9	0	0.0
Chiton	1	0.7	1	0.2	0	0.0
Epibionts or brought in incidentally						
Mollusca						
Turritellidae	0	0	1	<0.1	0	0
Helcion sp.	1	<0.1	1	<0.1	0	0
Nodilittorina spp.	1	0.1	4	0.3	1	0.2
Crustacea						
Cirripedia						
Shore barnacle	0	0	1	0.4	1	0.8
Coronula spp. (whale barnacle)	1	1.1	0	0	0	0

MNI, minimum number of individuals.

additions of a new class of food to the human diet before the introduction of domesticates at the end of the Pleistocene. Coastlines have few resources to attract hunter-gatherers if their diets do not include shellfish and/or fish. Once they do, coastlines become attractive for settlement and movement. It has been argued that shellfish exploitation was crucial to a potential early coastal route of modern humans out of Africa via the Red Sea coast[8], and marine adaptations made possible an early migration to Australia/New Guinea along a coastal corridor[26,27]. Our results show that the coastal adaptation was present in South Africa long before the postulated dates for these migrations (after 120 kyr ago). Shellfish may have been a critical food source to the survival of human populations when they were faced with depressed terrestrial productivity during glacial stages such as MIS6, a time when much of southern Africa was more arid[28] and populations were isolated and perhaps concentrated on now-submerged coastal platforms.

Shellfish can be a predictable food resource for humans[9] with substantial nutritional benefits[29]. Shellfish collecting is often associated with hunter-gatherer economic and social systems with greater complexity and reduced mobility[9], which are themselves excellent contexts for stimulating symbolic expression through material culture. The PP13B ochre sample has all the hallmarks of pigment for body-painting and perhaps colouring of other organic surfaces, and thus joins a patchy sample of Middle Pleistocene evidence for pigment use. By 164 kyr ago, there is preferential processing of the reddest pigments, as is present in more recent sites[16], showing the same pattern of pigment exploitation 40 kyr before its apparent fluorescence post 120 kyr ago. We have identified the earliest appearance of a dietary, technological and cultural package that included coastal occupation, bladelet technology, pigment use and dietary expansion to marine shellfish, and is dated to a time close to the biological emergence of modern humans.

METHODS SUMMARY

Owing to the small amount of preserved LC-MSA sediment, we conducted a limited excavation of a 1.5-m N–S section. We excavated within 50 cm × 50 cm quadrants within squares, named by their bearing: NE, NW, SE and SW. Excavations followed natural stratigraphic units (layers, features, and so on), and thus square-quadrant stratigraphic unit provenance designation is the minimum assigned to any find. Sediment volumes were measured during excavation, and bulk samples of sediment were taken from every unique stratigraphic unit. All observed finds were plotted directly to Total Station in three dimensions, whereas the rest were captured by nested 10-mm→3-mm→1.5-mm wet-sieving. Screened materials were dried, packed in plastic bags and transported to the field laboratory. All plotted finds were labelled with their specimen number in black India ink. Finds were then sorted in the laboratory and provided to the appropriate specialist for analysis (lithics, ochre and shellfish; see Author Contributions). Inclinations of finds were calculated from two shots on opposite ends of the finds. Lithics were analysed by a combination of typological, technological and metrical variables from a database of all plotted finds, and those from the 10-mm mesh screen. Ochre was studied under a 10–40× zoom microscope, and streak properties were ascertained by the production of a streak across white porcelain plates. Shellfish were identified by comparison to known modern specimens. The backgrounds of all photographic images of artefacts in Fig. 2 were removed. No portion of any artefact image was retouched or otherwise edited. Stratigraphic interpretations are derived from a combination of field-based macro-stratigraphic observations, computer analysis of mapped stratigraphic units, analyses of plotted find distributions, and micromorphology. Dating of the sediments is accomplished by OSL and U-series techniques.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 May; accepted 28 August 2007.

1. Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408,** 708–713 (2000).
2. McDougall, I., Brown, F. H. & Fleagle, J. G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433,** 733–736 (2005).
3. Henshilwood, C., d'Errico, F., Vanhaeren, M., van Niekerk, K. & Jacobs, Z. Middle Stone Age shell beads from South Africa. *Science* **304,** 404 (2004).
4. Vanhaeren, M. *et al.* Middle Paleolithic shell beads in Israel and Algeria. *Science* **312,** 1785–1788 (2006).
5. Petit, J. R. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399,** 429–436 (1999).
6. Marean, C. W. & Assefa, Z. in *African Archaeology* (ed. Stahl, A. B.) 93–129 (Blackwell, New York, 2005).
7. McBrearty, S. & Brooks, A. S. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J. Hum. Evol.* **39,** 453–563 (2000).
8. Walter, R. C. *et al.* Early human occupation of the Red Sea coast of Eritrea during the last interglacial. *Nature* **405,** 65–69 (2000).
9. Erlandson, J. M. The Archaeology of Aquatic Adaptations: Paradigms for a New Millennium. *J. Archaeol. Res.* **9,** 287–350 (2001).
10. Ambrose, S. H. Small things remembered: origins of early microlithic industries in sub-saharan Africa. *Archaeol. Pap. Am. Anthropol. Assoc.* **12,** 9–29 (2002).
11. Soriano, S., Villa, P. & Wadley, L. Blade technology and tool forms in the Middle Stone Age of South Africa: the Howiesons Poort and post-Howiesons Poort at Rose Cottage Cave. *J. Archaeol. Sci.* **34,** 681–703 (2007).
12. Villa, P., Delagnes, A. & Wadley, L. A late Middle Stone Age artifact assemblage from Sibudu (KwaZulu-Natal): comparisons with the European Middle Paleolithic. *J. Archaeol Sci.* **32,** 399–422 (2005).
13. Tryon, C. A. & McBrearty, S. Tephrostratigraphy and the Acheulian to Middle Stone Age transition in the Kapthurin Formation, Kenya. *J. Hum. Evol.* **42,** 211–236 (2002).
14. Hendey, Q. B. & Volman, T. P. Last interglacial sea levels and coastal caves in the Cape Province, South Africa. *Quat. Res.* **2,** 189–198 (1986).
15. Henshilwood, C. S., D'Errico, F., Marean, C. W., Milo, R. G. & Yates, R. J. An early bone tool industry from the Middle Stone Age, Blombos Cave, South Africa: implications for the origins of modern human behaviour, symbolism and language. *J. Hum. Evol.* **41,** 631–678 (2001).
16. Watts, I. Ochre in the Middle Stone Age of southern Africa: Ritualized display or hide preservative? *S. Afr. Archaeol. Bull.* **57,** 1–14 (2002).
17. Hovers, E. *et al.* An early case of color symbolism. *Curr. Anthropol.* **44,** 491–522 (2003).
18. Barham, L. S. Systematic pigment use in the Middle Pleistocene of South-Central Africa. *Curr. Anthropol.* **43,** 181–190 (2002).
19. Marean, C. W., Nilssen, P. J., Brown, K., Jerardino, A. & Stynder, D. Paleoanthropological investigations of Middle Stone Age sites at Pinnacle Point, Mossel Bay (South Africa): Archaeology and hominid remains from the 2000 Field Season. *Paleoanthropology* **2,** 14–83 (2004).
20. Waelbroeck, C. *et al.* Sea-level and deep water temperature changes derived from benthic foraminifera isotopic records. *Quat. Sci. Rev.* **21,** 295–305 (2002).
21. Bigalke, E. H. The exploitation of shellfish by coastal tribesmen of the Transkei. *Ann. Cape Prov. Mus. Nat. Hist.* **9,** 159–175 (1973).
22. Tryon, C. A., McBrearty, S. & Texier, P. J. Levallois lithic technology from the Kapthurin Formation, Kenya: Acheulian Origin and Middle Stone Age Diversity. *Afr. Archaeol. Rev.* **22,** 199–229 (2005).
23. Branch, G. *Two Oceans: A Guide to the Marine Life of Southern Africa* (D. Philip, Cape Town, 1994).

24. Newman, W. A. & Ross, A. Antarctic Cirripedia: Monographic Account Based on Specimens Collected Chiefly Under the United States Antarctic Research Program, 1962–1965. *Am. Geophys. Union. Antarct. Res. Ser.* **14,** 1–257 (1971).
25. Jerardino, A. & Parkington, J. New evidence for whales on archaeological sites in the south-western Cape. *S. Afr. J. Sci.* **89,** 6–7 (1993).
26. Kingdon, J. *Self-made man and his undoing* (Simon and Schuster, London, 1993).
27. Bulbeck, D. Where River Meets Sea: A Parsimonious Model for *Homo sapiens* Colonization of the Indian Ocean Rim and Sahul. *Curr. Anthropol.* **48,** 315–321 (2007).
28. Deacon, J. & Lancaster, N. *Late Quaternary Paleoenvironments of Southern Africa* (Clarendon Press, Oxford, 1988).
29. Broadhurst, C. L. *et al.* Brain-specific lipids from marine, lacustrine, or terrestrial food resources: potential impact on early African *Homo sapiens*. *Comp. Biochem. Phys. B.* **131,** 653–673 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the ISSR staff at ASU, the MAP staff for their assistance, the Dias Museum for field facilities, SAHRA and HWC for permits, and Waelbroeck for helping with sea level data. This research was funded by grants from the National Science Foundation (to C.W.M.) and the Hyde Family Foundation (to C.W.M.).

Author Contributions C.W.M. directed the excavations and is the project principal investigator. Authors made contributions in the following areas: M.B.-M., U-series dating; J.B., analysis of orientation and dip; E.F., three-dimensional GIS; P.G. and P.K., micromorphology and geology; A.I.R.H., geology and sediment magnetics; Z.J., OSL dating; A.J., shell analysis; T.M., E.T. and H.M.W., lithics; P.J.N., co-direction of the excavations; and I.W., ochre. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.W.M. (curtis.marean@asu.edu).

METHODS

Excavation methods. Site PP13B was excavated within an arbitrary three-dimensional coordinate system (grid) called the MAP grid. All horizontal coordinates reported in this paper are in that grid space. It is tied to the South African National Coordinate Reference System, and our elevation measurements are in orthometric height above sea level as defined by that coordinate system. Our horizontal coordinates can be transformed to the South African grid using a two-dimensional conformal coordinate transformation, and from there to latitude and longitude following the Gauss Kruger Conform Projection using the Hartebeesthoek 1994 datum used by South Africa. In the MAP coordinate system, a 1 m square is named by the planar coordinates of its south-west corner. Published descriptions of our excavation methods are available[30,31].

Photography methods. The incorrect light settings of the camera for the image in Fig. 2b were corrected using Nikon Capture Editor 4 White Balance, with the camera white balance reset to 'daylight'. This correction was applied to the entire image. Figure 2c is a cropped higher resolution image of the whole artefact shown in Fig. 2b.

Much of the section photography presented here is geo-referenced to the MAP grid, unless otherwise indicated. This is accomplished in the following way. Targets are attached to the section approximately every 50 cm and these are shot to the Total Station. The photographs are then brought into ESRI ArcGIS and rectified (rotated and stretched) to the grid using those targets and their coordinates. This process creates a photograph that is true to elevation and one dimension of grid space to within several centimetres of its true geometric space, allowing one to plot on it finds that have been shot to the Total Station. There is a slight degradation of the image during the rectification process. For this reason, we have provided unedited versions of the two stratigraphic photographs (Supplementary Figs 1 and 2).

The micromorphology photographs (Supplementary Figs 3 and 4) were taken with a Zeiss Axioplan 40 POL polarizing microscope and an attached digital camera (Canon Powershot G5). A microscope $\times$12.5 magnification was used and a $\times$4 camera zoom. Digital image settings were auto-adjusted (for exposure, shadows, brightness and contrast) with Photoshop CS2. This adjustment was made uniformly to the entire image. No portion of the photographs was retouched or otherwise edited.

Lithic analysis methods. In the laboratory, the plotted and 10-mm screened materials were not washed. Each complete and fragmentary artefact was measured in standard multiple dimensions with a large number of traits recorded in a relational database. A basic shape typology and Geneste's[32] technological typology was recorded. All cores were recorded using the MSA typology of Volman[33] and the Geneste technological typology. Raw materials were recorded on the basis of geologic classes, which were easy to visually discriminate (only 0.2%, $N = 4$ unidentified/other).

Sea level GIS model method. The multidimensional GIS databases discussed in the text and featured within the Supplementary Video 1 were created using ESRI ArcScene and ArcGlobe 9.2. ESRI ArcScene was used primarily to model site-based archaeological and geological features within PP13B. The PP13B cave model represents a collection of more than 800 points shot along the floor and walls of the cave with a reflectorless Total Station. These points were then used to create a TIN model with Raindrop Geomagic software. The LC-MSA stratigraphic units visible in the model are a combination of 2.5-dimensional polygons representing the upper and lower surfaces of each unit and multipatch polygon convex hulls.

ESRI ArcGlobe was used for regional sea level modelling throughout the Pinnacle Point/Mossel Bay area. Base terrestrial elevations are a combination of regional SRTM 90 m digital elevation data and 1 m light detection and ranging (LIDAR) data around Pinnacle Point. The Pinnacle Point cliff face is a TIN model created from Total Station point survey data (over 30,000 shots) using Raindrop Geomagic. Additional imagery within the database includes 30 m resolution Landsat ETM+ images (Regional) and sub-orbital 1 m aerial photography (Pinnacle Point). The bathymetric data are a combination of high-resolution 1 m side-scanning sound navigation and ranging (SONAR) around Pinnacle Point and hand-digitization from published regional bathymetry maps. Sea level data were created by clipping the bathymetry elevation model to the listed depths below current relative sea level[34] and then converting the data into *z*-aware 3D polygons. Tabular output of this model is presented in Supplementary Table 1.

U-series dating methods. Six flowstone samples were collected from the flowstone capping the LC-MSA by chipping with a hammer from the horizontal extent of the *in situ* flowstone. The entire flowstone as well as each sample was mapped by Total Station. The flowstone was then cored to produce the micromorphology sample. Details of the U-series laboratory methods are presented in Supplementary Information.

OSL dating methods. Nine sediment samples were collected from the LC-MSA in the North Area of Cave 13B (Supplementary Table 2). Five of the samples were collected from the LC-MSA Upper, of which two (46447 and 111400) were from the upper-dune sand and three (46467, 20720 and 111401) were from the lower-dune sand (see Supplementary Information for discussion of stratigraphy details). The other four sediment samples were collected from the LC-MSA archaeological sediments: sample 111402 from the LC-MSA Middle, and samples 111403, 111404 and 20721 from the LC-MSA Lower sediments. We also collected one additional sample (111406) from the lowermost unexcavated LC-MSA sediments cemented to the south cave wall. Details of the OSL laboratory methods and results are presented in Supplementary Information.

30. Marean, C. W., Nilssen, P. J., Brown, K., Jerardino, A. & Stynder, D. Paleoanthropological investigations of Middle Stone Age sites at Pinnacle Point, Mossel Bay (South Africa): Archaeology and hominid remains from the 2000 Field Season. *Paleoanthropology* **2,** 14–83 (2004).
31. Dibble, H., Marean, C. & McPherron, S. P. The use of barcodes in excavation projects. *The SAA Archaeological Record* **7,** 33–38 (2007).
32. Geneste, J. M. *Analyse Lithique d'Industries Mousteriennes Perigord: Une Approche Technologique du Comportement des Groupes Humains au Paleolithique Moyen*. PhD thesis, Univ. of Bordeaux (1985).
33. Volman, T. P. *The Middle Stone Age in the Southern Cape*. PhD thesis, Univ. of Chicago (1981).
34. Waelbroeck, C. *et al.* Sea-level and deep water temperature changes derived from benthic foraminifera isotopic records. *Quat. Sci. Rev.* **21,** 295–305 (2002).

LETTERS

Turnover of sex chromosomes induced by sexual conflict

G. S. van Doorn[1,2] & M. Kirkpatrick[2]

Sex-determination genes are among the most fluid features of the genome in many groups of animals[1,2]. In some taxa the master sex-determining gene moves frequently between chromosomes, whereas in other taxa different genes have been recruited to determine the sex of the zygotes. There is a well developed theory for the origin of stable and highly dimorphic sex chromosomes seen in groups such as the eutherian mammals[3]. In contrast, the evolutionary lability of genetic sex determination in other groups remains largely unexplained[1]. In this theoretical study, we show that an autosomal gene under sexually antagonistic selection can cause the spread of a new sex-determining gene linked to it. The mechanism can account for the origin of new sex-determining loci, the transposition of an ancestral sex-determining gene to an autosome, and the maintenance of multiple sex-determining factors in species that lack heteromorphic sex chromosomes.

Fish provide examples of the dynamic nature of genetic sex determination seen in some groups of animals[4]. At least four different chromosomes determine sex in different species of salmon[5], the master sex-determining gene can differ between congeneric species[6], and sex determination is polygenic in some fish species[7].

Several mechanisms have been suggested to explain the puzzling diversity of genetic sex-determination mechanisms. These include random genetic drift[1,8], pleiotropic selection favouring new sex-determining alleles[9,10], sex-ratio selection[11,12] and various kinds of transmission distortion[13]. Although each is plausible for certain cases, these mechanisms involve fairly special biological conditions (for example, small population size or fortuitous pleiotropy).

Here we suggest a mechanism that extends the theory on the origin of sex chromosomes[1,14] to explain the movement of male determination from an ancestral Y chromosome to an autosome that then invades to become a neo-Y chromosome. The underlying force driving the change is sexually antagonistic selection, which is thought to be widespread on both theoretical and empirical grounds[15].

The mechanism begins with an autosomal locus segregating for two alleles that have sexually antagonistic effects (that is, different relative fitnesses in males and females). Consider the consequences of a mutation nearby on the same chromosome that causes individuals to develop into males regardless of what sex chromosomes they carry. This mutation could occur in a gene involved in the sex-determination cascade, for example, or result from transposition of the male-determining factor from the Y chromosome to the autosome. A genetic association (linkage disequilibrium) will develop naturally between the new allele that makes zygotes male and the allele that makes them good at being male. If this combination of genes produces males that have higher fitness than those carrying the original Y, the neo-Y can spread, effectively hijacking sex determination from the original sex chromosomes.

This verbal argument raises a series of questions. For example, how will additional sexually antagonistic loci located on the original sex chromosomes affect the process? Will invasion of a neo-Y always cause the loss of the ancestral Y, or can both be maintained in a multifactorial sex-determination system?

To address these issues, we developed a formal population-genetic model consisting of four loci. The first two are sex-determination factors: locus *Y* is the ancestral master sex-determination gene located on the sex chromosomes, whereas the autosomal locus *y* carries a dominant masculinizing mutation. The remaining two loci each segregate for two alleles with sexually antagonistic effects. Locus *a* is on the same autosome as locus *y*, whereas locus *A* is on the ancestral sex chromosome with locus *Y*. Locus *A* is included to account for the effects of genes with sex-antagonistic effects that tend to accumulate on the sex chromosomes[16]. Our primary aim is to explain the lability of sex determination in groups without highly differentiated sex chromosomes. We therefore assume that the sex chromosomes are non-heteromorphic. Locus *A* is present on both X and Y chromosomes, and we allow for recombination between *A* and *Y*. The evolutionary dynamics of the model are described by a system of 255 equations. Although it is not possible to do a full analysis, we were able to derive an approximation that describes how the population evolves when either the new masculinizing mutation or the ancestral Y chromosome is rare. Details are given in the Supplementary Information, where we also support our results by exploring the consequences of alternative assumptions on the genetic properties of the new sex-determining allele (partial dominance, incomplete penetrance or recessiveness).

When the masculinizing mutation is rare, its frequency changes at the exponential rate:

$$\lambda = S_a L_a V_a - S_A L_A V_A \quad (1)$$

The mutation spreads if λ is positive, and is lost if it is negative. The first of the two terms on the right represents the effect of locus *a*, which is linked to the new masculinizing mutation and favours it to invade. The second term results from locus *A*, which is carried on the ancestral sex chromosome and inhibits invasion of the new mutation. This inhibition is a consequence of the linkage disequilibrium between the ancestral sex-determining factor and male-beneficial alleles at locus *A*. Males that carry the neo-Y also carry two ancestral X chromosomes. The ancestral X chromosomes are enriched for the sex-antagonistic allele that is beneficial to females. Normal males carry an ancestral Y chromosome, which, in contrast, is enriched for the male-beneficial allele. Neo-Y carriers thus suffer a fitness reduction, quantified exactly by the second term on the right-hand side of equation (1). Both this fitness reduction and the fitness gain resulting from the genetic association between the new masculinizing

[1]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA. [2]Section of Integrative Biology, University of Texas, 1 University Station C-0930, Austin, Texas 78712, USA.

factor at locus y and the male-beneficial allele at locus a can be decomposed in three contributing factors. The coefficients S_a and S_A represent the strength of sexually antagonistic selection acting on a and A, whereas $L_a = (1-r_a)/r_a$ and $L_A = (1-r_A)/r_A$ measure how closely linked those loci are to the sex-determining genes on their respective chromosomes (r_a and r_A are the recombination rates). The last elements of equation (1) are V_a and V_A, which measure the genetic variation at loci a and A. S and V depend on the allele frequencies at the sex-antagonistic loci, and their values can evolve. Full definitions for S and V are given in the Methods.

Equation (1) verifies the verbal argument: a masculinizing mutation can spread because of sexually antagonistic selection. The mutation's evolutionary advantage is strengthened by stronger sex-antagonistic selection and greater genetic variation at locus a, as well as tighter linkage between that gene and the new masculinizing factor at locus y. Conversely, sexually antagonistic selection acting on locus A on the sex chromosome favours the ancestral Y chromosome over the new mutation. Selection favours the Y chromosome that has the highest mean fitness, which in turn is determined by the pattern of sex-antagonistic selection and the amount of recombination.

What is the ultimate fate of a masculinizing mutation if it does invade? We can determine this fate by noting that equation (1) describes the dynamics of the ancestral Y chromosome when it is rare if we interchange indices A and a, and recalculate the values of S and V for the case that nearly all males carry the neo-Y. The simplest situation is when the sex-antagonistic genes are loosely linked to the sex determination loci (L_a, $L_A \ll 1$); in this case, the values of S and V change very little as the masculinizing mutation spreads (see Methods). Consequently, equation (1) implies that conditions that favour the new masculinizing mutation to spread when it is rare also favour the ancestral Y to be lost when it is rare. In short, if the masculinizing mutation increases when rare, it will spread to fixation. This process is exemplified by Fig. 1, which shows, for a particular set of parameters, predictions for the relative growth rates based on equation (1) together with corresponding simulation results. The agreement between the analytical approximation and the exact numerical simulations is generally as accurate as in Fig. 1b when selection is weak.

In the case illustrated by Fig. 1, sex determination is hijacked by the autosome from the ancestral sex chromosomes. The ancestral Y

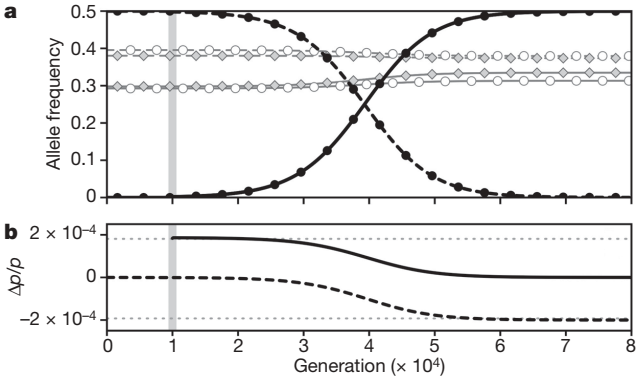


Figure 1 | Sex determination hijacked by an autosomal sex-determining factor. **a**, Black lines show the evolution of allele frequencies at sex-determination loci in males during a sex-chromosome switch (solid line, neo-Y; dashed line, ancestral Y). The frequencies of sex-antagonistic alleles change only slightly as the neo-Y spreads to fixation. Grey lines depict frequencies in females (open circles) and males (filled diamonds) at loci a (solid) and A (dashed). **b**, Equation (1) accurately predicts the asymptotic values (grey dotted lines) of the relative rates of increase of the neo-Y ($\Delta p_y/p_y$, solid black line) and the ancestral Y ($\Delta p_Y/p_Y$, dashed black line). The grey bar in panels **a** and **b** marks when the neo-Y first appeared by mutation. Parameters are: $s_A^F = 0.024$, $s_A^M = -0.026$, $s_a^F = -0.029$, $s_a^M = 0.025$, $h_A^F = h_a^M = 0.6$, $h_A^M = h_a^F = 0.4$, $r_A = 0.12$, $r_a = 0.08$.

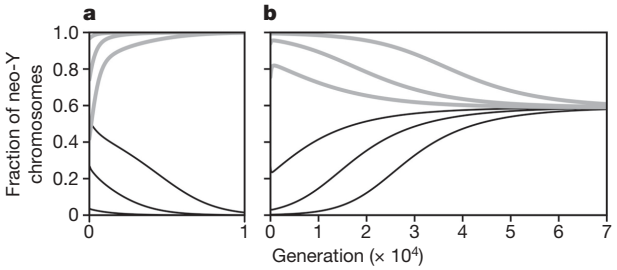


Figure 2 | Bistability and protected polymorphism of sex-determining factors. The two panels show examples of cases in which invasion of the neo-Y and its potential to spread to fixation do not coincide. **a**, Bistability: the neo-Y cannot invade a population in which sex is determined by the ancestral sex chromosomes (thin black lines depict runs with different initial frequencies of the neo-Y), but neither can the ancestral Y when the neo-Y is the established sex-determination factor (thick grey lines, for three different initial frequencies of the ancestral Y). Selection coefficients are: $s_A^F = -0.028$, $s_A^M = 0.017$, $s_a^F = -0.023$, $s_a^M = 0.027$. **b**, Protected polymorphism: the neo-Y can invade, but it cannot completely replace the ancestral Y, resulting in multifactorial sex determination (this is for $s_A^F = -0.027$, $s_A^M = 0.018$, $s_a^F = -0.028$, $s_a^M = 0.022$). Other parameters, for both panels, are: $h_A^F = 0.375$, $h_A^M = 0.625$, $h_a^F = 0.4$, $h_a^M = 0.6$, $r_A = 0.009$, $r_a = 0.012$.

disappears and the ancestral X becomes a new autosome. A neo-X and neo-Y are formed from the autosome that carries the masculinizing locus y. During this substitution YY males are not produced and so the potential deleterious effects of such genotypes do not affect the evolutionary process. Moreover, the substitution does not affect the sex ratio, which remains stable at 1:1 throughout.

More complex outcomes can occur when the sex-determining and sexually antagonistic loci are tightly linked (Figs 2 and 3). Here the dynamics of the sex-determination factors can induce considerable change in the genetic variances at the sexually antagonistic loci, such that invasion of the masculinizing mutation no longer implies loss of the ancestral Y. For some combinations of viability effects and linkage, both the ancestral Y and the new masculinizing mutation are lost when rare (Fig. 2a and region 3 in Fig. 3). The system is thus bistable: the population evolves to a single-factor sex-determination system governed by either locus Y or locus y, depending on the initial conditions. It is possible that random genetic drift could trigger

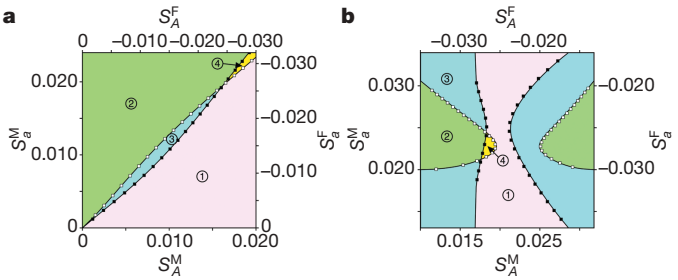


Figure 3 | Dependence of evolutionary outcomes on the selection coefficients. Systematically varying the values of the selection coefficients, we delineated four regions in parameter space that correspond to qualitatively different evolutionary outcomes of the model. In region 1 the ancestral Y is stable against invasion and further spread of the neo-Y. In region 2 the neo-Y can invade and replace the ancestral Y (as in Fig. 1). Regions 3 and 4 demarcate the selective regimes that give rise to bistability (as in Fig. 2a) or stable multifactorial sex determination (Fig. 2b), respectively. The boundaries between the regions were calculated from equation (1) (black lines) and by means of exact numerical simulations (open squares mark the invasion boundary of the neo-Y; filled squares mark its fixation boundary). In **a**, we varied the magnitude of sex-antagonistic fitness effects while keeping the ratio s_i^F/s_i^M ($i = a$ or A) constant. In **b**, the difference $s_i^M - s_i^F$ was fixed and we varied the average of the selection coefficients for males and females. Other parameters are as in Fig. 2.

transitions between these two equilibria. For other selection regimes, both the ancestral Y chromosome and the masculinizing mutation will increase when rare (Fig. 2b and region 4 in Fig. 3). The result is a protected polymorphism at both sex-determining loci, and the population evolves a two-factor sex-determination system. To our knowledge, sex-antagonistic selection is the only mechanism known that can produce a nuclear sex-determination system that can show stable multifactorial inheritance (see ref. 17) or bistability.

All else being equal, bistability and protected polymorphism occur when the intensities of sexually antagonistic selection at the sex-linked and autosomal loci are of comparable magnitude. This can be seen in Fig. 3a, in which the regions 3 and 4 extend over the diagonal. Away from the diagonal, one of the sex factors is associated with significantly stronger sex-antagonistic fitness effects, and that factor replaces the other. Figure 3b provides further insights in the population-genetic mechanisms responsible for bistability and protected polymorphism. Bistability is prominent in the corner regions of Fig. 3b, where the fitness effects of the sex-antagonistic alleles are strongly biased towards one sex or the other. In such cases, much less genetic variation can be maintained at autosomal loci than at sex-linked loci. Whichever sex factor is rare is thus linked to a sex-antagonistic locus that harbours little genetic variation, whereas the established sex factor is linked to a more variable sex-antagonistic locus. This causes an intrinsic disadvantage of rarity resulting in bistability. The opposite effect acts on sex-antagonistic alleles that are nearly neutral on average, and this explains why the region of protected polymorphism is located centrally in Fig. 3b. Sexually antagonistic alleles with equal but opposite fitness effects are maintained at a frequency close to one-half at autosomal loci, but tend to go to fixation at sex-linked loci, especially when linkage is tight. Genetic variation at the sex-linked locus, expressed as an average of X- and Y-linked variation (see Methods), will thus be smaller than the genetic variation at the autosomal sex-antagonistic locus. The result is an inherent advantage of rare sex factors allowing for the maintenance of multiple sex-determination alleles.

Three factors inhibit the hijacking process and might account for the great evolutionary stability of sex chromosomes in groups such as mammals and birds[18]. The first is the presence of genes essential for male fertility or viability that are located on the ancestral Y chromosome and that are absent from the ancestral X. Unless the neo-Y resulted from a major translocation containing the male-determining factor and the essential genes from the ancestral Y chromosome, such genes would absolutely prevent the invasion of the new masculinizing factor. A second inhibiting factor is the evolution of dosage compensation in genes that are close to the sex-determining locus, and the third brake on the process is produced by sex-antagonistic genes on the ancestral sex chromosomes (represented by the second term of equation (1)). Long-term evolution of the sex chromosomes typically results in the accumulation of sex-antagonistic polymorphisms[16], the reduction of recombination rates[1,3] and divergence of the X and Y chromosomes in the vicinity of the master sex-determining gene. As the sex chromosomes progressively differentiate, these factors make the conditions for hijacking more restrictive (by increasing S_A and L_A in equation (1)), enhancing the evolutionary stability of the established sex-determination system. In contrast, the evolution of sex-limited expression of sex-antagonistic genes on the ancestral sex chromosomes makes these chromosomes more vulnerable to the invasion of new sex-determining factors. Sex-limited expression reduces the sexual conflict, or may even fully resolve it, leading to a loss of polymorphism at the ancestral sex chromosomes. The long-term stability or lability of sex-determination may thus depend on a balance between sexual conflict, the evolution of gene regulation and structural evolution of the sex chromosomes[18].

What is the scope for the mechanism described here? The two essential ingredients are sexually antagonistic polymorphisms and new sex-determining loci on autosomes. Polymorphism at sexually antagonistic loci can be maintained by constant selection pressures, as we assumed in this study, but only for a restricted range of parameters, particularly at autosomal loci. Yet, the mechanism we portray here can also operate if some other evolutionary force maintains the sexually antagonistic polymorphism, for example, frequency-dependent selection, migration or mutation. Alternatively, even transient polymorphisms could trigger the hijack mechanism, providing that the total fitness variation at sexually antagonistic loci generated by transient polymorphisms is sufficiently large at any point in time. Recent data from expression studies reveal that a remarkable fraction—between 15% and 70%—of genes has sexually dimorphic expression in a variety of organisms[19–21]. Any segregating gene among those with sex-specific effects could participate in the hijack process. Given the ubiquity of sexually dimorphic expression, we do not expect sex-chromosome switches to be precluded by a lack of variation at sexually antagonistic loci, even if we are ignorant about the mechanisms that support the high levels of polymorphism found in nature.

The second ingredient, sex-determining mutations on autosomes, may also be quite common. Our model applies equally to transposition of an existing master sex-determining gene as to mutations at other loci that result in sex determination. Both processes are known. In humans, for example, there are many autosomal mutations that reverse sex (see http://www.ncbi.nlm.nih.gov/omim/). Translocation of a master male-determining gene to an autosome has been suggested in several groups of animals (for example, flies[22,23] and salmonid fishes[5]). Thus, in some taxa there may be a sufficient flux of mutations that satisfy equation (1) to explain the observed turnover of sex chromosomes. Another possibility is that an inversion can, by chance, capture a masculinizing allele and a sex-antagonistic gene, instantly increasing the linkage between the two (the term L_a in equation (1)) and therefore triggering a hijack.

In the discussion above, males are the heterogametic sex (that is, the sex determination system is XY). The mechanism also applies to female heterogamety (ZW sex determination, as in birds and butterflies), in which case a dominant feminizing mutation on an autosome hijacks sex determination from the ancestral sex chromosomes. The model does not address heterogamety switches, however, in which there is an evolutionary transition between XY and ZW sex determination. We expect that heterogamety switches, which are known from several groups of vertebrates[2], might also be driven by sexually antagonistic selection. The evolutionary process involved, however, is more complex because YY (or WW) individuals are produced.

A prediction from our model is that recently derived sex-determining regions will be associated with genes that are targets of sexually antagonistic selection. Observations consistent with this prediction are that sexually selected colour genes are closely linked to the sex-determining genes in poeciliid[7] and cichlid[24] fishes. This is a weak test of the hypothesis, however, because the sexually antagonistic genes may have accumulated after the new sex chromosomes were established rather than driving the process. A more stringent test would be to look for sexually antagonistic genes in very young sex chromosomes, and in the homologous autosomal regions of closely related species that have not undergone the hijacking. Promising systems for these investigations include the medaka[6] and the three-spined stickleback[25].

Sexually antagonistic selection is thought to result most often from behavioural strategies shaped by sexual selection, through either male–male competition or female choice[15]. Although it has long been known that genes contribute importantly to differences in behaviour between individuals within a species, the model presented here suggests that the arrow of causality can also point in the opposite direction. Behaviour may drive the evolution of the genome, as well as the converse.

METHODS SUMMARY

The relative viabilities of the (0,0), (0,1) and (1,1) genotypes in females are 1: $1+h_i^F s_i^F$: $1+s_i^F$ for locus i (= A or a), where s_i^F and h_i^F represent selection and

dominance coefficients, respectively. The notation for viabilities in males is analogous, but with F (female) replaced by M (male). We assume that loci *A* and *a* have independent (multiplicative) effects on fitness and that mating is random.

We distinguish between the frequency of allele 1 at locus *A* on the ancestral X chromosome, denoted p_A^X, and its frequency on the ancestral Y chromosome, denoted p_A^Y. The frequency of allele 1 at locus *a* on chromosomes carrying the masculinizing mutation at locus *y* is denoted p_a^y, and its frequency averaged over all chromosomes is $\bar{p}_a$.

The factors S_A and S_a appearing in equation (1) measure the effects that sexually antagonistic selection on loci *A* and *a* have on the masculinizing mutation at locus *y*. These terms, which are derived in the Supplementary Information, are defined as:

$$S_a = \frac{1}{2} s_a^M[\bar{p}_a + h_a^M(1-2\bar{p}_a)]\{s_a^M[\bar{p}_a + h_a^M(1-2\bar{p}_a)] - s_a^F[\bar{p}_a + h_a^F(1-2\bar{p}_a)]\}$$

$$S_A = \frac{4\{s_A^M[p_A^X + h_A^M(1-2p_A^X)]\}^2\{2s_A^F[p_A^X + h_A^F(1-2p_A^X)] + s_A^M[p_A^Y + h_A^M(1-2p_A^Y)]\}}{2s_A^F[p_A^X + h_A^F(1-2p_A^X)] + s_A^M[p_A^Y + h_A^M(1-2p_A^Y)] - 3s_A^M[p_A^X + h_A^M(1-2p_A^X)]}$$

The terms in equation (1) that represent genetic variation at those loci are: $V_a = p_a^y(1-p_a^y)$, $V_A = \frac{1}{4}[3p_A^X(1-p_A^X) + p_A^Y(1-p_A^Y)]$

For our analyses, we evaluated these expressions using the equilibrium allele frequencies at loci *A* and *a* before the masculinizing mutation appears[16]. Those frequencies depend only weakly on the frequency of the neo-Y chromosome when linkage is weak (L_A, $L_a \ll 1$) (Fig. 1a), a fact that can be used to show that if the masculinizing mutation at locus *y* is favoured when rare then the ancestral Y chromosome will be lost.

The analytical results presented in equation (1) and the Supplementary Information were checked by means of numerical simulations based on a full set of recursions for the genotype frequencies that did not involve the approximations used in the analytical treatment. Results of these simulations are shown in Figs 1–3.

Received 16 July; accepted 17 August 2007.

1. Bull, J. J. *Evolution of Sex Determining Mechanisms* (Benjamin/Cummings, Reading, Massachusetts, 1983).
2. Ezaz, T., Stiglec, R., Veyrunes, F., &. Marshall Graves, J. A. Relationships between vertebrate ZW and XY sex chromosome systems. *Curr. Biol.* **16,** R736–R743 (2006).
3. Charlesworth, B. The evolution of sex chromosomes. *Science* **251,** 1030–1033 (1991).
4. Mank, J. E., Promislow, D. E. L. & Avise, J. C. Evolution of alternative sex-determining mechanisms in teleost fishes. *Biol. J. Linn. Soc.* **87,** 83–93 (2006).
5. Woram, R. A. *et al.* Comparative genome analysis of the primary sex-determining locus in salmonid fishes. *Genome Res.* **13,** 272–280 (2003).
6. Matsuda, M. Sex determination in the teleost medaka, *Oryzias latipes*. *Annu. Rev. Genet.* **39,** 293–307 (2005).
7. Kallman, K. D. in *Evolutionary Genetics of Fishes* (ed. Turner, B. J.) 95–171 (Plenum, New York, 1984).
8. Vuilleumier, S., Lande, R., Van Alphen, J. J. M. & Seehausen, O. Invasion and fixation of sex-reversal genes. *J. Evol. Biol.* **20,** 913–920 (2007).
9. Seehausen, O., Van Alphen, J. J. M. & Lande, R. Color polymorphism and sex ratio distortion in a cichlid fish as an incipient stage in sympatric speciation by sexual selection. *Ecol. Lett.* **2,** 367–378 (1999).
10. Pomiankowski, A., Nothiger, R. & Wilkins, A. The evolution of the *Drosophila* sex-determination pathway. *Genetics* **166,** 1761–1773 (2004).
11. Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Rev. Genet.* **5,** 288–298 (2004).
12. Kozielska, M., Pen, I., Beukeboom, L. W. & Weissing, F. J. Sex ratio selection and multi-factorial sex determination in the housefly: a dynamic model. *J. Evol. Biol.* **19,** 879–888 (2006).
13. Werren, J. H. & Beukeboom, L. W. Sex determination, sex ratios, and genetic conflict. *Annu. Rev. Ecol. Syst.* **29,** 233–261 (1998).
14. Rice, W. R. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. *Evolution* **41,** 911–914 (1987).
15. Arnqvist, G. & Rowe, L. *Sexual Conflict* (Princeton Univ. Press, Princeton, 2005).
16. Rice, W. R. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38,** 735–742 (1984).
17. Rice, W. R. On the instability of polygenic sex determination: the effect of sex-specific selection. *Evolution* **40,** 633–639 (1986).
18. Marín, I. & Baker, B. S. The evolutionary dynamics of sex determination. *Science* **281,** 1990–1994 (1998).
19. Oliver, B. & Parisi, M. Battle of the Xs. *Bioessays* **26,** 543–548 (2004).
20. Gnad, F. & Parsch, J. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics* **22,** 2577–2579 (2006).
21. Yang, X. *et al.* Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* **16,** 995–1004 (2006).
22. Traut, W. Sex determination in the fly *Megaselia scalaris*, a model system for primary steps of sex chromosome evolution. *Genetics* **136,** 1097–1104 (1994).
23. Carvalho, A. B. & Clark, A. G. Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila* Y. *Science* **307,** 108–110 (2005).
24. Lande, R., Seehausen, O. & Van Alphen, J. J. M. Mechanisms of rapid sympatric speciation by sex reversal and sexual selection in cichlid fish. *Genetica* **112–113,** 435–443 (2001).
25. Peichel, C. L. *et al.* The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. *Curr. Biol.* **14,** 1416–1424 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. J. Bull, J. Mank and J. Ranz for discussions. G.S.vD. was supported by a Rubicon grant from the Netherlands Organisation for Scientific Research (NWO).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.S.vD. (vandoorn@santafe.edu).

LETTERS

Genome-wide detection and characterization of positive selection in human populations

Pardis C. Sabeti[1]*, Patrick Varilly[1]*, Ben Fry[1], Jason Lohmueller[1], Elizabeth Hostetter[1], Chris Cotsapas[1,2], Xiaohui Xie[1], Elizabeth H. Byrne[1], Steven A. McCarroll[1,2], Rachelle Gaudet[3], Stephen F. Schaffner[1], Eric S. Lander[1,4,5,6] & The International HapMap Consortium†

With the advent of dense maps of human genetic variation, it is now possible to detect positive natural selection across the human genome. Here we report an analysis of over 3 million polymorphisms from the International HapMap Project Phase 2 (HapMap2)[1]. We used 'long-range haplotype' methods, which were developed to identify alleles segregating in a population that have undergone recent selection[2], and we also developed new methods that are based on cross-population comparisons to discover alleles that have swept to near-fixation within a population. The analysis reveals more than 300 strong candidate regions. Focusing on the strongest 22 regions, we develop a heuristic for scrutinizing these regions to identify candidate targets of selection. In a complementary analysis, we identify 26 non-synonymous, coding, single nucleotide polymorphisms showing regional evidence of positive selection. Examination of these candidates highlights three cases in which two genes in a common biological process have apparently undergone positive selection in the same population: *LARGE* and *DMD*, both related to infection by the Lassa virus[3], in West Africa; *SLC24A5* and *SLC45A2*, both involved in skin pigmentation[4,5], in Europe; and *EDAR* and *EDA2R*, both involved in development of hair follicles[6], in Asia.

An increasing amount of information about genetic variation, together with new analytical methods, is making it possible to explore the recent evolutionary history of the human population. The first phase of the International Haplotype Map, including ~1 million single nucleotide polymorphisms (SNPs)[7], allowed preliminary examination of natural selection in humans. Now, with the publication of the Phase 2 map (HapMap2)[1] in a companion paper, over 3 million SNPs have been genotyped in 420 chromosomes from three continents (120 European (CEU), 120 African (YRI) and 180 Asian from Japan and China (JPT + CHB)).

In our analysis of HapMap2, we first implemented two widely used tests that detect recent positive selection by finding common alleles carried on unusually long haplotypes[2]. The two, the Long-Range Haplotype (LRH)[8] and the integrated Haplotype Score (iHS)[9] tests, rely on the principle that, under positive selection, an allele may rise to high frequency rapidly enough that long-range association with nearby polymorphisms—the long-range haplotype[8]—will not have time to be eliminated by recombination. These tests control for local variation in recombination rates by comparing long haplotypes to other alleles at the same locus. As a result, they lose power as selected alleles approach fixation (100% frequency), because there are then few alternative alleles in the population (Supplementary Fig. 2 and Supplementary Tables 1–2).

We next developed, evaluated and applied a new test, Cross Population Extended Haplotype Homozygosity (XP-EHH), to detect selective sweeps in which the selected allele has approached or achieved fixation in one population but remains polymorphic in the human population as a whole (Methods, and Supplementary Fig. 2 and Supplementary Tables 3–6). Related methods have recently also been described[10–12].

Our analysis of recent positive selection, using the three methods, reveals more than 300 candidate regions[1](Supplementary Fig. 3 and Supplementary Table 7), 22 of which are above a threshold such that no similar events were found in 10 Gb of simulated neutrally evolving sequence (Methods). We focused on these 22 strongest signals (Table 1), which include two well-established cases, *SLC24A5* and *LCT*[2,5,13], and 20 other regions with signals of similar strength.

The challenge is to sift through genetic variation in the candidate regions to identify the variants that were the targets of selection. Our candidate regions are large (mean length, 815 kb; maximum length, 3.5 Mb) and often contain multiple genes (median, 4; maximum, 15). A typical region harbours ~400–4,000 common SNPs (minor allele frequency >5%), of which roughly three-quarters are represented in current SNP databases and half were genotyped as part of HapMap2 (Supplementary Table 8).

We developed three criteria to help highlight potential targets of selection (Supplementary Fig. 1): (1) selected alleles detectable by our tests are likely to be derived (newly arisen), because long-haplotype tests have little power to detect selection on standing (pre-existing) variation[14]; we therefore focused on derived alleles, as identified by comparison to primate outgroups; (2) selected alleles are likely to be highly differentiated between populations, because recent selection is probably a local environmental adaptation[2]; we thus looked for alleles common in only the population(s) under selection; (3) selected alleles must have biological effects. On the basis of current knowledge, we therefore focused on non-synonymous coding SNPs and SNPs in evolutionarily conserved sequences. These criteria are intended as heuristics, not absolute requirements. Some targets of selection may not satisfy them, and some will not be in current SNP databases. Nonetheless, with ~50% of common SNPs in these populations genotyped in HapMap2, a search for causal variants is timely.

We applied the criteria to the regions containing *SLC24A5* and *LCT*, each of which already has a strong candidate gene, mutation and trait. At *SLC24A5*, the 600 kb region contains 914 genotyped

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA. [2]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [4]Department of Biology, MIT, Cambridge, Massachusetts 02139, USA. [5]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. [6]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.
†Lists of participants and affiliations appear at the end of the paper.

Table 1 | The twenty-two strongest candidates for natural selection

Region	Chr:position (MB, HG17)	Selected population	Long Haplotype Test	Size (Mb)	Total SNPs with Long Haplotype Signal	Subset of SNPs that fulfil criteria 1	Subset of SNPs that fulfil criteria 1 and 2	Subset of SNPs that fulfil criteria 1, 2 and 3	Genes at or near SNPs that fulfil all three criteria
1	chr1:166	CHB + JPT	LRH, iHS	0.4	92	39	30	2	*BLZF1, SLC19A2*
2	chr2:72.6	CHB + JPT	XP-EHH	0.8	732	250	0	0	
3	chr2:108.7	CHB + JPT	LRH, iHS, XP-EHH	1.0	972	265	7	1	*EDAR*
4	chr2:136.1	CEU	LRH, iHS, XP-EHH	2.4	1,213	282	24	3	*RAB3GAP1, R3HDM1, LCT*
5	chr2:177.9	CEU,CHB + JPT	LRH, iHS, XP-EHH	1.2	1,388	399	79	9	*PDE11A*
6	chr4:33.9	CEU,YRI, CHB + JPT	LRH, iHS	1.7	413	161	33	0	
7	chr4:42	CHB + JPT	LRH, iHS, XP-EHH	0.3	249	94	65	6	*SLC30A9*
8	chr4:159	CHB + JPT	LRH, iHS, XP-EHH	0.3	233	67	34	1	
9	chr10:3	CEU	LRH, iHS, XP-EHH	0.3	179	63	16	1	
10	chr10:22.7	CEU, CHB + JPT	XP-EHH	0.3	254	93	0	0	
11	chr10:55.7	CHB+JPT	LRH, iHS, XP-EHH	0.4	735	221	5	2	*PCDH15*
12	chr12:78.3	YRI	LRH, iHS	0.8	151	91	25	0	
13	chr15:46.4	CEU	XP-EHH	0.6	867	233	5	1	*SLC24A5*
14	chr15:61.8	CHB + JPT	XP-EHH	0.2	252	73	40	6	*HERC1*
15	chr16:64.3	CHB + JPT	XP-EHH	0.4	484	137	2	0	
16	chr16:74.3	CHB + JPT, YRI	LRH, iHS	0.6	55	35	28	3	*CHST5, ADAT1, KARS*
17	chr17:53.3	CHB + JPT	XP-EHH	0.2	143	41	0	0	
18	chr17:56.4	CEU	XP-EHH	0.4	290	98	26	3	*BCAS3*
19	chr19:43.5	YRI	LRH, iHS, XP-EHH	0.3	83	30	0	0	
20	chr22:32.5	YRI	LRH	0.4	318	188	35	3	*LARGE*
21	chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
22	chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	
		Total SNPs		16.74	9,166	2,898	480	41	

Twenty-two regions were identified at a high threshold for significance (Methods), based on the LRH, iHS and/or XP-EHH test. Within these regions, we examined SNPs with the best evidence of being the target of selection on the basis of having a long haplotype signal, and by fulfilling three criteria: (1) being a high-frequency derived allele; (2) being differentiated between populations and common only in the selected population; and (3) being identified as functional by current annotation. Several candidate polymorphisms arise from the analysis including well-known *LCT* and *SLC24A5* (ref. 2), as well as intriguing new candidates.

SNPs. Applying filters progressively (Table 1 and Fig. 1a–d), we found that 867 SNPs are associated with the long-haplotype signal, of which 233 are high-frequency derived alleles, of which 12 are highly differentiated between populations, and of which only 5 are common in Europe and rare in Asia and Africa. Among these five SNPs, there is only one implicated as functional by current knowledge; it has the strongest signal of positive selection and encodes the A111T polymorphism associated with pigment differences in

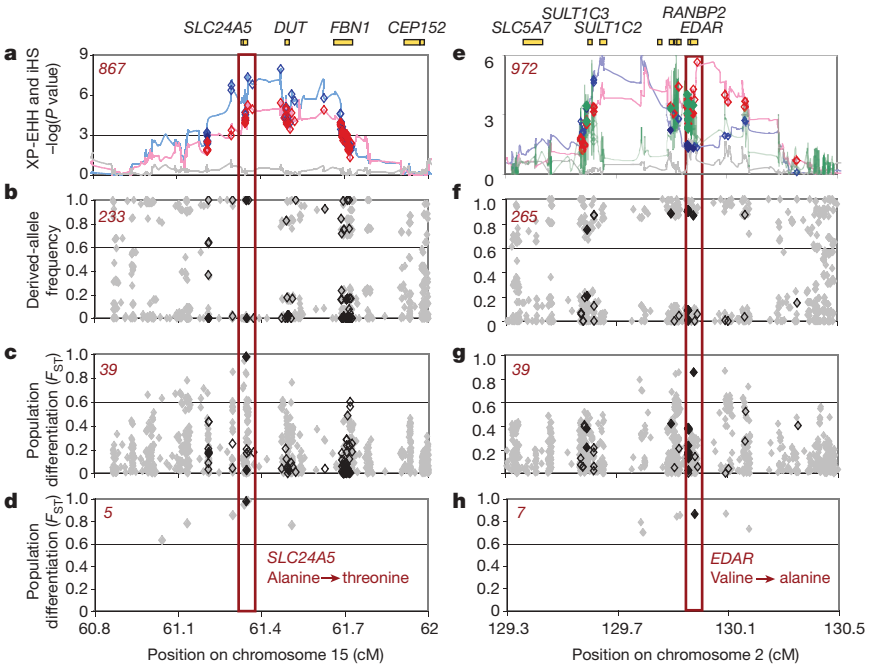


Figure 1 | Localizing *SLC24A5* and *EDAR* signals of selection. **a–d**, *SLC24A5*. **a**, Strong evidence for positive selection in CEU samples at a chromosome 15 locus: XP-EHH between CEU and JPT + CHB (blue), CEU and YRI (red), and YRI and JPT + CHB (grey). SNPs are classified as having low probability (bordered diamonds) and high probability (filled diamonds) potential for function. SNPs were filtered to identify likely targets of selection on the basis of the frequency of derived alleles (**b**), differences between populations (**c**) and differences between populations for high-frequency derived alleles (less than 20% in non-selected populations) (**d**). The number of SNPs that passed each filter is given in the top left corner in red. The threonine to alanine candidate polymorphism in *SLC24A5* is the clear outlier. **e–h**, *EDAR*. **e**, Similar evidence for positive selection in JPT + CHB at a chromosome 2 locus: XP-EHH between CEU and JPT + CHB (blue), between YRI and JPT + CHB (red), and between CEU and YRI (grey); iHS in JPT + CHB (green). A valine to alanine polymorphism in *EDAR* passes all filters: the frequency of derived alleles (**f**), differences between populations (**g**) and differences between populations for high-frequency derived alleles (less than 20% in non-selected populations) (**h**). Three other functional changes, a D→E change in *SULT1C2* and two SNPs associated with *RANBP2* expression (Methods), have also become common in the selected population.

humans and thought to be the target of positive selection[5]. Our criteria thus uniquely identify the expected allele.

At the *LCT* locus, we found similar degrees of filtration. Within the 2.4 Mb selective sweep, 24 polymorphisms fulfil the first two criteria (Table 1, and Supplementary Fig. 4), with the polymorphism thought to confer adult persistence of lactase among them. However, this SNP was only identified as functional after extensive study of the *LCT* gene[15]. Thus *LCT* shows both the utility and the limits of the heuristics.

Given the encouraging results for *SLC24A5* and *LCT*, we performed a similar analysis on all 22 candidate regions (Table 1). Filtering the 9,166 SNPs associated with the long-haplotype signal, we found that 480 satisfied the first two criteria. We identified 41 out of the 480 SNPs (0.2% of all SNPs genotyped in the regions) as possibly functional on the basis of a newly compiled database of polymorphisms in known coding elements, evolutionarily conserved elements and regulatory elements (Methods; B.F., unpublished), together containing $\sim$ 5.5% of all known SNPs.

Eight of the forty-one SNPs encode non-synonymous changes (Table 1 and Supplementary Table 9). Apart from the well-known case of *SLC24A5*, they are found in *EDAR*, *PCDH15*, *ADAT1*, *KARS*, *HERC1*, *SLC30A9* and *BLFZ1*. The remaining 33 potentially functional SNPs lie within conserved transcription factor motifs, introns, UTRs and other non-coding regions.

To identify additional candidates, we reversed the process by taking non-synonymous coding SNPs with highly differentiated high-frequency derived alleles; these SNPs comprise a tiny fraction of all SNPs and have a higher a priori probability of being targets of selection. Of the 15,816 non-synonymous SNPs in HapMap2, 281 (Supplementary Table 10) have both a high derived-allele frequency (frequency >50%) and clear differentiation between populations (F_{ST} is in the top 0.5 percentile). We examined these 281 SNPs to identify those embedded within long-range haplotypes[16], and identified 26 putative cases of positive selection. These include the eight non-synonymous SNPs identified in the genome-wide analysis above.

Interestingly, analysis of the top regions and the non-synonymous SNPs together revealed three cases of two genes in the same pathway both having strong evidence of selection in a single population.

In the European sample, there is strong evidence for two genes already shown to be associated with skin pigment differences among humans. The first is *SLC24A5*, described above. We further examined the global distribution (Fig. 2) and the predicted effect on protein activity of the *SLC24A5* A111T polymorphism (Supplementary Fig. 5, 6). The second, *SLC45A2*, has an important role in pigmentation in zebrafish, mouse and horse[4]. An L374F substitution in *SLC45A2* is at 100% frequency in the European sample, but absent in the Asian and African samples. A recent association study has shown that the Phe-encoding allele is correlated with fair skin and non-black hair in Europeans[4]. Together, the data support *SLC45A2* as a target of positive selection in Europe[10,17].

In the African sample (Yoruba in Ibadan, Nigeria), there is evidence of selection for two genes with well-documented biological links to the Lassa fever virus. The strongest signal in the genome, on the basis of the LRH test, resides within a 400 kb region that lies entirely within the gene *LARGE*. The LARGE protein is a glycosylase that post-translationally modifies α-dystroglycan, the cellular receptor for Lassa fever virus (as well as other arenaviruses), and the modification has been shown to be critical for virus binding[3]. The virus name is derived from Lassa, Nigeria, where the disease is endemic, with 21% of the population showing signs of exposure[18]. We also noted that the *DMD* locus is on our larger candidate list of regions, with the signal of selection again in the Yoruba sample. *DMD* encodes a cytosolic adaptor protein that binds to α-dystroglycan and is critical for its function. We hypothesize that Lassa fever created selective pressure at *LARGE* and *DMD*[12]. This hypothesis can be tested by correlating the geographical distribution of the selected haplotype with endemicity of the Lassa virus, studying infection of genotyped cells *in vitro*, and searching for an association between the selected haplotype and clinical outcomes in infected patients.

In the Asian samples, we found evidence of selection for non-synonymous polymorphisms in two genes in the ectodysplasin (EDA) pathway, which is involved in development of hair, teeth and exocrine glands[6]. The genes are *EDAR* and *EDA2R*, which encode the key receptors for the ligands EDA A1 and EDA A2, respectively. Notably, the EDA signalling pathway has been shown to be under positive selection for loss of scales in multiple distinct populations of freshwater stickleback fish[19]. A mutation encoding a V370A substitution in *EDAR* is near fixation in Asia and absent in Europe and Africa (Fig. 1e–h). An R57K substitution in *EDA2R* has derived-allele frequencies of 100% in Asia, 70% in Europe and 0% in Africa.

The *EDAR* polymorphism is notable because it is highly differentiated between the Asian and other continental populations (the 3rd most differentiated among 15,816 non-synonymous SNPs), and also within Asian populations (in the top 1% of SNPs differentiated between the Japanese and Chinese HapMap samples). Genotyping of the *EDAR* polymorphism in the CEPH (Centre d'Etudie du Polymorphisme Humain) global diversity panel[20] shows that it is at high but varying frequency throughout Asia and the Americas (for example, 100% in Pima Indians and in parts of China, and 73% in Japan) (Fig. 2, and Supplementary Fig. 7). Studying populations like the Japanese, in which the allele is still segregating, may provide clues to its biological significance.

EDAR has a central role in generation of the primary hair follicle pattern, and mutations in *EDAR* cause hypohidrotic ectodermal

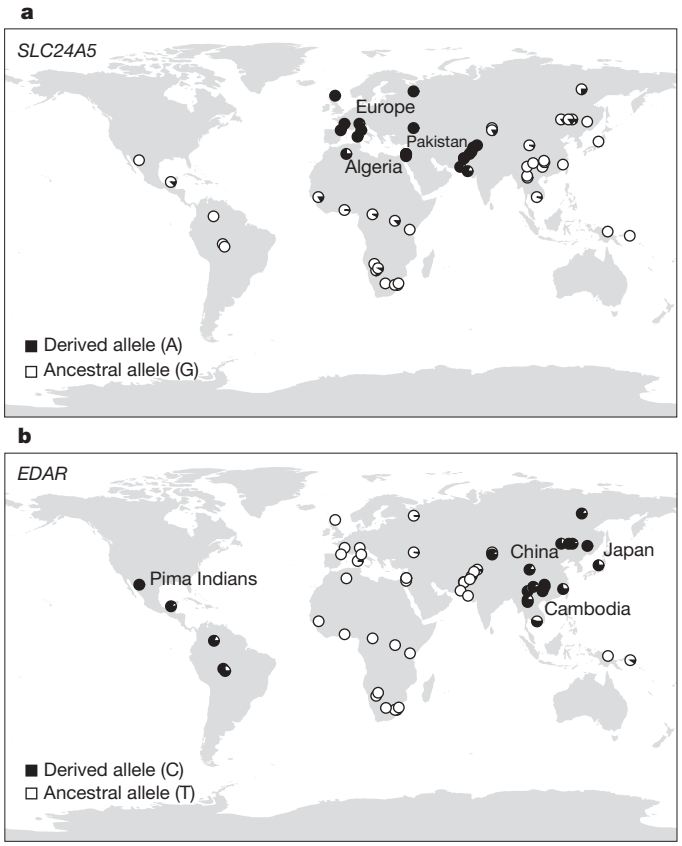


Figure 2 | Global distribution of *SLC24A5* A111T and *EDAR* V370A. Worldwide allele-frequency distributions for candidate polymorphisms with the strongest evidence for selection[20]. **a**, *SLC24A5* A111T is common in Europe, Northern Africa and Pakistan, but rare or absent elsewhere. **b**, *EDAR* V370A is common in Asia and the Americas, but absent in Europe and Africa.

dysplasia (HED) in humans and mice, characterized by defects in the development of hair, teeth and exocrine glands[6]. The V370A polymorphism, proposed to be the target of selection, lies within *EDAR*'s highly conserved death domain (Supplementary Fig. 8), the location of the majority of *EDAR* polymorphisms causing HED[21]. Our structural modelling predicts that the polymorphism lies within the binding site of the domain (Fig. 3).

Our analysis only scratches the surface of the recent selective history of the human genome. The results indicate that individual candidates may coalesce into pathways that reveal traits under selection, analogous to the alleles of multiple genes (for example, *HBB*, *G6PD* and *DARC*) that arose and spread in Africa and other tropical populations as a result of the partial protection they confer against malaria[2,12]. Such endeavours will be enhanced by continuing development of analytical methods to localize signals in candidate regions, generation of expanded data sets, advances in comparative genomics to define coding and regulatory regions, and biological follow-up of promising candidates. True understanding of the role of adaptive evolution will require collaboration across multiple disciplines, including molecular and structural biology, medical and population genetics, and history and anthropology.

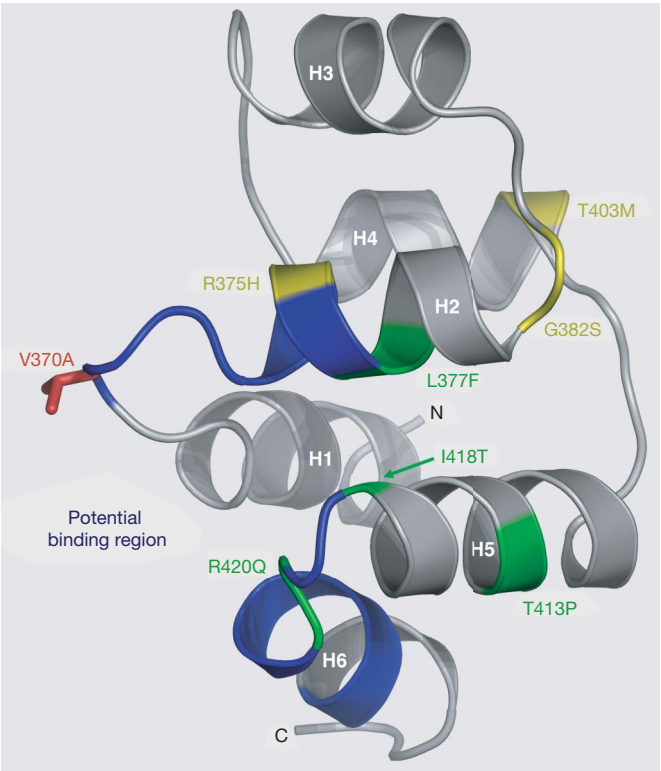


Figure 3 | Structural model of the EDAR death domain. Ribbon representation of a homology model of the EDAR death domain (DD), based on the alignment of the EDAR DD amino acid sequence (EDAR residues 356–431), with multiple known DD structures. The helices are labelled H1 to H6. Residues in blue (the H1–H2 and H5–H6 loops, residues 370–376 and 419–425, respectively) correspond to the homologous residues in Tube that interact with Pelle in the Tube-DD–Pelle-DD structure[24]. These EDAR-DD residues therefore form a potential region of interaction with a DD-containing EDAR-interacting protein, such as EDARADD. The V370A polymorphic residue (red) is located prominently within this potential binding region in the H1–H2 loop. Seven of the thirteen known mis-sense mutations in EDAR that lead to hypohidrotic ectodermal dysplasia (HED) in humans are located in the EDAR-DD: the only four mutations in EDAR that lead to the dominant transmission of HED (green) and three recessive mutations (yellow)[21]. Four of these mutations, R375H, L377F, R420Q and I418T are located in the vicinity of the predicted interaction interface.

METHODS SUMMARY

Genotyping data. Phase 2 of the International Haplotype Map (HapMap2) (www.hapmap.org) contains 3.1 million SNPs genotyped in 420 chromosomes in 3 continental populations (120 European (CEU), 120 African (YRI) and 180 Asian (JPT+CHB))[1]. We further genotyped our top HapMap2 functional candidates in the HGDR-CEPH Human Genome Diversity Cell Line Panel[20].

LRH, iHS and XP-EHH tests. The Long-Range Haplotype (LRH), integrated Haplotype Score (iHS) and Cross Population EHH (XP-EHH) tests detect alleles that have risen to high frequency rapidly enough that long-range association with nearby polymorphisms—the long-range haplotype—has not been eroded by recombination; haplotype length is measured by the EHH[8,9]. The first two tests detect partial selective sweeps, whereas XP-EHH detects selected alleles that have risen to near fixation in one but not all populations. To evaluate the tests, we simulated genomic data for each HapMap population in a range of demographic scenarios—under neutral evolution and twenty scenarios of positive selection—developing the program Sweep (www.broad.mit.edu/mpg/sweep) for analysis. For our top candidates by the three tests, we tested for haplotype-specific recombination rates and copy-number polymorphisms, possible confounders.

Localization. We calculated F_{ST} and derived-allele frequency for all SNPs within the top candidate regions. We developed a database for those regions to annotate all potentially functional DNA changes (B.F., unpublished), including non-synonymous variants, variants disrupting predicted functional motifs, variants within regions of conservation in mammals and variants previously associated with human phenotypic differences, as well as synonymous, intronic and untranslated region variants.

Structural model. We generated a homology model of the EDAR death domain (DD) from available DD structures using Modeller 9v1 (ref. 22). The distribution of conserved residues, built using ConSurf[23] with an EDAR sequence alignment from 22 species, shows a bias to the protein core in helices H1, H2 and H5, supporting our model.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 August; accepted 13 September 2007.

1. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* doi:10.1038/nature06258 (this issue).
2. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312,** 1614–1620 (2006).
3. Kunz, S. *et al.* Posttranslational modification of α-dystroglycan, the cellular receptor for arenaviruses, by the glycosyltransferase LARGE is critical for virus binding. *J. Virol.* **79,** 14282–14296 (2005).
4. Graf, J., Hodgson, R. & van Daal, A. Single nucleotide polymorphisms in the *MATP* gene are associated with normal human pigmentation variation. *Hum. Mutat.* **25,** 278–284 (2005).
5. Lamason, R. L. *et al.* SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310,** 1782–1786 (2005).
6. Botchkarev, V. A. & Fessing, M. Y. Edar signaling in the control of hair follicle development. *J. Investig. Dermatol. Symp. Proc.* **10,** 247–251 (2005).
7. The International Haplotype Map Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
8. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
9. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).
10. Kimura, R., Fujimoto, A., Tokunaga, K. & Ohashi, J. A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS ONE* **2,** e286 (2007).
11. Tang, K., Thornton, K. R. & Stoneking, M. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* **5,** e171 (2007).
12. Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3,** e90 (2007).
13. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74,** 1111–1120 (2004).
14. Teshima, K. M., Coop, G. & Przeworski, M. How reliable are empirical genomic scans for selective sweeps? **16,** 702–712 *Genome Res.* (2006).
15. Kuokkanen, M. *et al.* Transcriptional regulation of the lactase–phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. *Gut* **52,** 647–652 (2003).
16. Miller, R. G. *Simultaneous statistical inference* XVI 299 (Springer, New York, 1981).
17. Soejima, M., Tachida, H., Ishida, T., Sano, A. & Koda, Y. Evidence for recent positive selection at the human *AIM1* locus in a European population. *Mol. Biol. Evol.* **23,** 179–188 (2006).
18. Richmond, J. K. & Baglole, D. J. Lassa fever: epidemiology, clinical features, and social consequences. *Br. Med. J.* **327,** 1271–1275 (2003).
19. Colosimo, P. F. *et al.* Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307,** 1928–1933 (2005).

20. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298,** 2381–2385 (2002).
21. Chassaing, N., Bourthoumieu, S., Cossee, M., Calvas, P. & Vincent, M. C. Mutations in *EDAR* account for one-quarter of non-*ED1*-related hypohidrotic ectodermal dysplasia. *Hum. Mutat.* **27,** 255–259 (2006).
22. Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).
23. Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33,** W299–W302 (2005).
24. Xiao, T., Towb, P., Wasserman, S. A. & Sprang, S. R. Three-dimensional structure of a complex between the death domains of Pelle and Tube. *Cell* **99,** 545–555 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements P.C.S. is funded by a Burroughs Wellcome Career Award in the Biomedical Sciences and has been funded by the Damon Runyon Cancer Fellowship and the L'Oreal for Women in Science Award. We thank A. Schier, B. Voight, R. Roberts, M. Kreiger, A. Abzhanov, D. Degusta, M. Burnette, E. Lieberman, M. Daly, D. Altshuler, D. Reich, D. Lieberman and I. Woods for helpful discussions on our analysis and results. We also thank L. Ziaugra, D. Tabbaa and T. Rachupka for experimental assistance. This work was funded in part by grants from the National Human Genome Research Institute (to E.S.L.) and from the Broad Institute of MIT and Harvard.

Author Contributions P.C.S., P.V., B.F. and E.S.L. initiated the project. P.V., B.F. and P.C.S. developed key software. P.C.S., P.V., B.F., S.F.S., J.L., E.H., C.C., X.X., E.B., S.A.McC. and R.G. performed analysis. P.C.S., E.B. and E.H. performed experiments. P.C.S., E.S.L., P.V. and S.F.S. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.C.S. (pardis@broad.mit.edu).

The International HapMap Consortium (Participants are arranged by institution and then alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.)

Genotyping centres: Perlegen Sciences Kelly A. Frazer (Principal Investigator)[1], Dennis G. Ballinger[2], David R. Cox[2], David A. Hinds[2], Laura L. Stuve[2]; **Baylor College of Medicine and ParAllele BioScience** Richard A. Gibbs (Principal Investigator)[3], John W. Belmont[3], Andrew Boudreau[4], Paul Hardenbol[5], Suzanne M. Leal[3], Shiran Pasternak[6], David A. Wheeler[3], Thomas D. Willis[4], Fuli Yu[7]; **Beijing Genomics Institute** Huanming Yang (Principal Investigator)[8], Changqing Zeng (Principal Investigator)[8], Yang Gao[8], Haoran Hu[8], Weitao Hu[8], Chaohua Li[8], Wei Lin[8], Siqi Liu[8], Hao Pan[8], Xiaoli Tang[8], Jian Wang[8], Wei Wang[8], Jun Yu[8], Bo Zhang[8], Qingrun Zhang[8], Hongbin Zhao[8], Hui Zhao[8], Jun Zhou[8]; **Broad Institute of Harvard and Massachusetts Institute of Technology** Stacey B. Gabriel (Project Leader)[7], Rachel Barry[7], Brendan Blumenstiel[7], Amy Camargo[7], Matthew Defelice[7], Maura Faggart[7], Mary Goyette[7], Supriya Gupta[7], Jamie Moore[7], Huy Nguyen[7], Robert C. Onofrio[7], Melissa Parkin[7], Jessica Roy[7], Erich Stahl[7], Ellen Winchester[7], Liuda Ziaugra[7], David Altshuler (Principal Investigator)[7,9]; **Chinese National Human Genome Center at Beijing** Yan Shen (Principal Investigator)[10], Zhijian Yao[10]; **Chinese National Human Genome Center at Shanghai** Wei Huang (Principal Investigator)[11], Xun Chu[11], Yungang He[11], Li Jin[12], Yangfan Liu[11], Yayun Shen[11], Weiwei Sun[11], Haifeng Wang[11], Yi Wang[11], Ying Wang[11], Xiaoyan Xiong[11], Liang Xu[11]; **Chinese University of Hong Kong** Mary M. Y. Waye (Principal Investigator)[13], Stephen K. W. Tsui[13]; **Hong Kong University of Science and Technology** Hong Xue (Principal Investigator)[14], J. Tze-Fei Wong[14]; **Illumina** Luana M. Galver (Project Leader)[15], Jian-Bing Fan[15], Kevin Gunderson[15], Sarah S. Murray[1], Arnold R. Oliphant[16], Mark S. Chee (Principal Investigator)[17]; **McGill University and Génome Québec Innovation Centre** Alexandre Montpetit (Project Leader)[18], Fanny Chagnon[18], Vincent Ferretti[18], Martin Leboeuf[18], Jean-François Olivier[4], Michael S. Phillips[18], Stéphanie Roumy[15], Clémentine Sallée[19], Andrei Verner[18], Thomas J. Hudson (Principal Investigator)[20]; **University of California at San Francisco and Washington University** Pui-Yan Kwok (Principal Investigator)[21], Dongmei Cai[21], Daniel C. Koboldt[22], Raymond D. Miller[22], Ludmila Pawlikowska[21], Patricia Taillon-Miller[22], Ming Xiao[21]; **University of Hong Kong** Lap-Chee Tsui (Principal Investigator)[23], William Mak[23], You Qiang Song[23], Paul K. H. Tam[23]; **University of Tokyo and RIKEN** Yusuke Nakamura (Principal Investigator)[24,25], Takahisa Kawaguchi[25], Takuya Kitamoto[25], Takashi Morizono[25], Atsushi Nagashima[25], Yozo Ohnishi[25], Akihiro Sekine[25], Toshihiro Tanaka[25], Tatsuhiko Tsunoda[25]; **Wellcome Trust Sanger Institute** Panos Deloukas (Project Leader)[26], Christine P. Bird[26], Marcos Delgado[26], Emmanouil T. Dermitzakis[26], Rhian Gwilliam[26], Sarah Hunt[26], Jonathan Morrison[27], Don Powell[26], Barbara E. Stranger[26], Pamela Whittaker[26], David R. Bentley (Principal Investigator)[28]

Analysis groups: Broad Institute Mark J. Daly (Project Leader)[7,9], Paul I. W. de Bakker[7,9], Jeff Barrett[7,9], Yves R. Chretien[7], Julian Maller[7,9], Steve McCarroll[7,9], Nick Patterson[7], Itsik Pe'er[29], Alkes Price[7], Shaun Purcell[9], Daniel J. Richter[7], Pardis Sabeti[7], Richa Saxena[7,9], Stephen F. Schaffner[7], Pak C. Sham[23], Patrick Varilly[7], David Altshuler (Principal Investigator)[7,9]; **Cold Spring Harbor Laboratory** Lincoln D. Stein (Principal Investigator)[6], Lalitha Krishnan[6], Albert Vernon Smith[6], Marcela K. Tello-Ruiz[6], Gudmundur A. Thorisson[30]; **Johns Hopkins University School of Medicine** Aravinda Chakravarti (Principal Investigator)[31], Peter E. Chen[31], David J. Cutler[31], Carl S. Kashuk[31], Shin Lin[31]; **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[32], Weihua Guan[32], Yun Li[32], Heather M. Munro[33], Zhaohui Steve Qin[32], Daryl J. Thomas[34]; **University of Oxford** Gilean McVean (Project Leader)[35], Adam Auton[35], Leonardo Bottolo[35], Niall Cardin[35], Susana Eyheramendy[35], Colin Freeman[35], Jonathan Marchini[35], Simon Myers[35], Chris Spencer[7], Matthew Stephens[36], Peter Donnelly (Principal Investigator)[35]; **University of Oxford, Wellcome Trust Centre for Human Genetics** Lon R. Cardon (Principal Investigator)[37], Geraldine Clarke[38], David M. Evans[38], Andrew P. Morris[38], Bruce S. Weir[39]; **RIKEN** Tatsuhiko Tsunoda (Principal Investigator)[25], Todd A. Johnson[25]; **US National Institutes of Health** James C. Mullikin[40]; **US National Institutes of Health National Center for Biotechnology Information** Stephen T. Sherry[41], Michael Feolo[41], Andrew Skol[42]

Community engagement/public consultation and sample collection groups: Beijing Normal University and Beijing Genomics Institute Houcan Zhang[43], Changqing Zeng[8], Hui Zhao[8]; **Health Sciences University of Hokkaido, Eubios Ethics Institute, and Shinshu University** Ichiro Matsuda (Principal Investigator)[44], Yoshimitsu Fukushima[45], Darryl R. Macer[46], Eiko Suda[47]; **Howard University and University of Ibadan** Charles N. Rotimi (Principal Investigator)[48], Clement A. Adebamowo[49], Ike Ajayi[49], Toyin Aniagwu[49], Patricia A. Marshall[50], Chibuzor Nkwodimmah[49], Charmaine D. M. Royal[48]; **University of Utah** Mark F. Leppert (Principal Investigator)[51], Missy Dixon[51], Andy Peiffer[51]

Ethical, legal and social issues: Chinese Academy of Social Sciences Renzong Qiu[52]; **Genetic Interest Group** Alastair Kent[53]; **Kyoto University** Kazuto Kato[54]; **Nagasaki University** Norio Niikawa[55]; **University of Ibadan School of Medicine** Isaac F. Adewole[49]; **University of Montréal** Bartha M. Knoppers[19]; **University of Oklahoma** Morris W. Foster[56]; **Vanderbilt University** Ellen Wright Clayton[57]; **Wellcome Trust** Jessica Watkin[58]

SNP discovery: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[3], John W. Belmont[3], Donna Muzny[3], Lynne Nazareth[3], Erica Sodergren[3], George M. Weinstock[3], David A. Wheeler[3], Imtaz Yakub[3]; **Broad Institute of Harvard and Massachusetts Institute of Technology** Stacey B. Gabriel (Project Leader)[7], Robert C. Onofrio[7], Daniel J. Richter[7], Liuda Ziaugra[7], Bruce W. Birren[7], Mark J. Daly[7,9], David Altshuler (Principal Investigator)[7,9]; **Washington University** Richard K. Wilson (Principal Investigator)[59], Lucinda L. Fulton[59]; **Wellcome Trust Sanger Institute** Jane Rogers (Principal Investigator)[26], John Burton[26], Nigel P. Carter[26], Christopher M. Clee[26], Mark Griffiths[26], Matthew C. Jones[26], Kirsten McLay[26], Robert W. Plumb[26], Mark T. Ross[26], Sarah K. Sims[26], David L. Willey[26]

Scientific management: Chinese Academy of Sciences Zhu Chen[60], Hua Han[60], Le Kang[60]; **Genome Canada** Martin Godbout[61], John C. Wallenburg[62]; **Génome Québec** Paul L'Archevêque[63], Guy Bellemare[63]; **Japanese Ministry of Education, Culture, Sports, Science and Technology** Koji Saeki[64]; **Ministry of Science and Technology of the People's Republic of China** Hongguang Wang[65], Daochang An[65], Hongbo Fu[65], Qing Li[65], Zhen Wang[65]; **The Human Genetic Resource Administration of China** Renwu Wang[66]; **The SNP Consortium** Arthur L. Holden[15]; **US National Institutes of Health** Lisa D. Brooks[67], Jean E. McEwen[67], Mark S. Guyer[67], Vivian Ota Wang[67,68], Jane L. Peterson[67], Michael Shi[69], Jack Spiegel[70], Lawrence M. Sung[71], Lynn F. Zacharia[67], Francis S. Collins[72]; **Wellcome Trust** Karen Kennedy[61], Ruth Jamieson[58], John Stewart[58]

[1]The Scripps Research Institute, 10550 North Torrey Pines Road MEM275, La Jolla, California 92037, USA. [2]Perlegen Sciences, 2021 Stierlin Court, Mountain View, California 94043, USA. [3]Baylor College of Medicine, Human Genome Sequencing Center, Department of Molecular and Human Genetics, 1 Baylor Plaza, Houston, Texas 77030, USA. [4]Affymetrix, 3420 Central Expressway, Santa Clara, California 95051, USA. [5]Pacific Biosciences, 1505 Adams Drive, Menlo Park, California 94025, USA. [6]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [7]The Broad Institute of Harvard and Massachusetts Institute of Technology, 1 Kendall Square, Cambridge, Massachusetts 02139, USA. [8]Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 100300, China. [9]Massachusetts General Hospital and Harvard Medical School, Simches Research Center, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [10]Chinese National Human Genome Center at Beijing, 3-707 N. Yongchang Road, Beijing Economic-Technological Development Area, Beijing 100176, China. [11]Chinese National Human Genome Center at Shanghai, 250 Bi Bo Road, Shanghai 201203, China. [12]Fudan University and CAS-MPG Partner Institute for Computational Biology, School of Life Sciences, SIBS, CAS, Shanghai, 201203, China. [13]The Chinese University of Hong Kong, Department of Biochemistry, The Croucher Laboratory for Human Genetics, 6/F Mong Man Wai Building, Shatin, Hong Kong. [14]Hong Kong University of Science and Technology, Department of Biochemistry and Applied Genomics Center, Clear Water Bay, Knowloon, Hong Kong. [15]Illumina, 9885 Towne Centre Drive, San Diego, California 92121, USA. [16]Complete Genomics, 658 North Pastoria Avenue, Sunnyvale, California 94085, USA. [17]Prognosys Biosciences, 4215 Sorrento Valley Boulevard, Suite 105, San Diego, California 92121, USA. [18]McGill University and Génome Québec Innovation Centre, 740 Dr Penfield Avenue, Montréal, Québec H3A 1A4, Canada. [19]University of Montréal, The Public Law Research Centre

(CRDP), PO Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada. [20]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 500, Toronto, Ontario, M5G 1L7, Canada. [21]University of California, San Francisco, Cardiovascular Research Institute, 513 Parnassus Avenue, Box 0793, San Francisco, California 94143, USA. [22]Washington University School of Medicine, Department of Genetics, 660 S. Euclid Avenue, Box 8232, St Louis, Missouri 63110, USA. [23]University of Hong Kong, Genome Research Centre, 6/F, Laboratory Block, 21 Sassoon Road, Pokfulam, Hong Kong. [24]University of Tokyo, Institute of Medical Science, 4-6-1 Sirokanedai, Minatoku, Tokyo 108-8639, Japan. [25]RIKEN SNP Research Center, 1-7-22 Suehiro-cho, Tsurumi-ku Yokohama, Kanagawa 230-0045, Japan. [26]Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [27]University of Cambridge, Department of Oncology, Cambridge CB1 8RN, UK. [28]Solexa, Chesterford Research Park, Little Chesterford, nr Saffron Walden, Essex CB10 1XL, UK. [29]Columbia University, 500 West 120th Street, New York, New York 10027, USA. [30]University of Leicester, Department of Genetics, Leicester LE1 7RH, UK. [31]Johns Hopkins University School of Medicine, McKusick-Nathans Institute of Genetic Medicine, Broadway Research Building, Suite 579, 733 N. Broadway, Baltimore, Maryland 21205, USA. [32]University of Michigan, Center for Statistical Genetics, Department of Biostatistics, 1420 Washington Heights, Ann Arbor, Michigan 48109, USA. [33]International Epidemiology Institute, 1455 Research Boulevard, Suite 550, Rockville, Maryland 20850, USA. [34]Center for Biomolecular Science and Engineering, Engineering 2, Suite 501, Mail Stop CBSE/ITI, UC Santa Cruz, Santa Cruz, California 95064, USA. [35]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK. [36]University of Chicago, Department of Statistics, 5734 S. University Avenue, Eckhart Hall, Room 126, Chicago, Illinois 60637, USA. [37]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. [38]University of Oxford/Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [39]University of Washington Department of Biostatistics, Box 357232, Seattle, Washington 98195, USA. [40]US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA. [41]US National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. [42]University of Chicago, Department of Medicine, Section of Genetic Medicine, 5801 South Ellis, Chicago, Illinois 60637, USA. [43]Beijing Normal University, 19 Xinjiekouwai Street, Beijing 100875, China. [44]Health Sciences University of Hokkaido, Ishikari Tobetsu Machi 1757, Hokkaido 061-0293, Japan. [45]Shinshu University School of Medicine, Department of Medical Genetics, Matsumoto 390-8621, Japan. [46]United Nations Educational, Scientific and Cultural Organization (UNESCO Bangkok), 920 Sukhumwit Road, Prakanong, Bangkok 10110, Thailand. [47]University of Tsukuba, Eubios Ethics Institute, PO Box 125, Tsukuba Science City 305-8691, Japan. [48]Howard University, National Human Genome Center, 2216 6th Street, NW, Washington, District of Columbia 20059, USA. [49]University of Ibadan College of Medicine, Ibadan, Oyo State, Nigeria. [50]Case Western Reserve University School of Medicine, Department of Bioethics, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. [51]University of Utah, Eccles Institute of Human Genetics, Department of Human Genetics, 15 North 2030 East, Salt Lake City, Utah 84112, USA. [52]Chinese Academy of Social Sciences, Institute of Philosophy/Center for Applied Ethics, 2121, Building 9, Caoqiao Xinyuan 3 Qu, Beijing 100067, China. [53]Genetic Interest Group, 4D Leroy House, 436 Essex Road, London N130P, UK. [54]Kyoto University, Institute for Research in Humanities and Graduate School of Biostudies, Ushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan. [55]Nagasaki University Graduate School of Biomedical Sciences, Department of Human Genetics, Sakamoto 1-12-4, Nagasaki 852-8523, Japan. [56]University of Oklahoma, Department of Anthropology, 455 W. Lindsey Street, Norman, Oklahoma 73019, USA. [57]Vanderbilt University, Center for Genetics and Health Policy, 507 Light Hall, Nashville, Tennessee 37232, USA. [58]Wellcome Trust, 215 Euston Road, London NW1 2BE, UK. [59]Washington University School of Medicine, Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA. [60]Chinese Academy of Sciences, 52 Sanlihe Road, Beijing 100864, China. [61]Genome Canada, 150 Metcalfe Street, Suite 2100, Ottawa, Ontario K2P 1P1, Canada. [62]McGill University, Office of Technology Transfer, 3550 University Street, Montréal, Québec H3A 2A7, Canada. [63]Génome Québec, 630, boulevard René-Lévesque Ouest, Montréal, Québec H3B 1S6, Canada. [64]Ministry of Education, Culture, Sports, Science, and Technology, 3-2-2 Kasumigaseki, Chiyodaku, Tokyo 100-8959, Japan. [65]Ministry of Science and Technology of the People's Republic of China, 15 B. Fuxing Road, Beijing 100862, China. [66]The Human Genetic Resource Administration of China, b7, Zaojunmiao, Haidian District, Beijing 100081, China. [67]US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA. [68]US National Institutes of Health, Office of Behavioral and Social Science Research, 31 Center Drive, Bethesda, Maryland 20892, USA. [69]Novartis Pharmaceuticals Corporation, Biomarker Development, One Health Plaza, East Hanover, New Jersey 07936, USA. [70]US National Institutes of Health, Office of Technology Transfer, 6011 Executive Boulevard, Rockville, Maryland 20852, USA. [71]University of Maryland School of Law, 500 W. Baltimore Street, Baltimore, Maryland 21201, USA. [72]US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.

METHODS

Genotyping data. The chromosomes examined in HapMap 2 were phased by the consortium using PHASE[25].

The HGDR-CEPH Human Genome Diversity Cell Line Panel[20] consists of 1,051 individuals from 51 populations across the world. We obtained DNA for the panel from the Foundation Jean Dausset (CEPH) and genotyped our top functional candidates for selection in the panel.

LRH, iHS, and XP-EHH tests. The Long-Range Haplotype (LRH) and the integrated Haplotype Score (iHS) tests have been previously described[8,9] and our methods are given in Supplementary Methods.

EHH between two SNPs, A and B, is defined as the probability that two randomly chosen chromosomes are homozygous at all SNPs between A and B, inclusive[8]; it is usually calculated using a sample of chromosomes from a single population. Explicitly, if the N chromosomes in a sample form G homozygous groups, with each group i having n_i elements, EHH is defined as

$$\mathrm{EHH} = \frac{\sum_{i=1}^{G} \binom{n_i}{2}}{\binom{N}{2}}$$

The XP-EHH test detects selective sweeps in which the selected allele has risen to high frequency or fixation in one population, but remains polymorphic in the human population as a whole; for this purpose it is more powerful than either iHS or LRH (Supplementary Fig. 2 and Supplementary Tables 3–6). XP-EHH uses cross-population comparison of haplotype lengths to control for local variation in recombination rates. Such cross-population comparison is complicated by the fact that haplotype lengths also depend on population history, such as bottlenecks and expansions[26]. The XP-EHH test normalizes for genome-wide differences in haplotype length between populations.

We define the XP-EHH test with respect to two populations, A and B, a given core SNP and a given direction (centromere distal or proximal). EHH is calculated for all SNPs in population A between the core SNP and X, and the value integrated with respect to genetic distance, with the result defined as I_A. I_B is defined analogously for population B. The statistic $\ln(I_A/I_B)$ is then calculated; an unusually positive value suggests selection in population A, a negative value selection in B. For identifying outliers, the log-ratio is normalized to have zero mean and unit variance. Details are given in Supplementary Methods.

We developed a computer program, Sweep, to implement these tests (LRH, iHS and XP-EHH) for positive selection, (Supplementary Methods; www.broad.mit.edu/mpg/sweep). In identifying the 22 strongest candidate regions, we considered regions with signals in at least two of five tests (LRH, iHS and XP-EHH in the three pairwise comparisons among the three populations), as well as those that had the strongest signal for each individual test. With this threshold we found no events in 10 Gb of simulated neutrally evolving sequence. For the top candidates by the three tests, we have taken additional steps to rule out the effects of recombination rate variation and copy number polymorphisms (Supplementary Methods).

Simulations and power calculations. We simulated the evolution of 1 MB sections of 120 chromosomes from each of the three continental HapMap populations, using a previously validated demographic model[27], under neutrality and under twenty scenarios of positive selection. We studied the effects of demography by further simulating recent bottlenecks with a range of intensity. Details of simulations and power calculations are given in Supplementary Methods.

Functional annotation. We developed an annotation database for our candidate regions to identify all DNA changes with potential functional consequence (B.F., unpublished). We first examined candidates most likely to be functional, including non-synonymous mutations, variants that disrupt predicted functional motifs (transcription factor motifs in conserved regions up to 10-kb 5′ of known genes and miRNA binding-site motifs in conserved 3′ untranslated regions of known genes), and variations reported to be associated with human phenotypic differences. For the last category, we identified variations associated with a clinical state (for example, malaria resistance) by a review of the published literature and those associated with changes to gene expression in lymphoblastoid cell lines from the HapMap individuals. The annotation included insertion/deletion mutations of all sizes. We also examined candidates with lower probability of being functional, including synonymous, intronic and untranslated variations and those that occur within regions of conservation in mammalian species. These methods are described in greater detail in Supplementary Methods.

Structural model of EDAR's death domain. We generated a homology model for EDAR's death domain (DD) using six solved DD structures: p75 NGFR-DD, RAIDD-DD, Pelle-DD, FADD-DD, Fas-DD and IRAK4-DD[24,28–32]. We aligned the corresponding protein sequences using SALIGN[33]. We then added the amino acid sequence of EDAR's DD (residues 356–431) to this structural alignment using Modeller 9v1 (ref. 22). The resulting alignment was used as the input to Modeller 9v1 to build ten EDAR-DD structure models, and the best model was selected based on the Objective Function Score. Owing to the high DOPE scores in the H1–H2 loop we performed a loop refinement using Modeller9v1, significantly reducing the energy of this region. We further evaluated the model by examining the distribution of conserved residues using ConSurf[23] with an alignment of EDAR-DD sequences from 22 species. We observed a bias of conserved residues to the protein core in H1, H2 and H5, which supports our EDAR-DD model. To identify potential binding regions of EDAR-DD, we used LSQMAN[34] to superimpose the model to the Tube-DD–Pelle-DD complex structure[24]. The H1–H2 and H5–H6 loops of the EDAR-DD correspond to Tube residues interacting with Pelle, and H2–H3 and H4–H5 loops to Pelle residues interacting with Tube. We focused our analysis on the residues corresponding to the interacting region in Tube because our EDAR-DD model is most similar to Tube. Figures were generated with PyMOL[35].

Other analysis. Description of methods for calculating F_{ST}, derived-allele frequency, alignment of the SLC24 amino acids, species alignments, conservation graphs, and estimation of the fraction of SNPs genotyped in HapMap2 and identified in dbSNP, are given in Supplementary Methods.

25. Stephens, M., Smith, N. J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68,** 978–989 (2001).
26. Crawford, D. C. *et al.* Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genet.* **36,** 700–706 (2004).
27. Schaffner, S. F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15,** 1576–1583 (2005).
28. Berglund, H. *et al.* The three-dimensional solution structure and dynamic properties of the human FADD death domain. *J. Mol. Biol.* **302,** 171–188 (2000).
29. Huang, B., Eberstadt, M., Olejniczak, E. T., Meadows, R. P. & Fesik, S. W. NMR structure and mutagenesis of the Fas (APO-1/CD95) death domain. *Nature* **384,** 638–641 (1996).
30. Lasker, M. V., Gajjar, M. M. & Nair, S. K. Cutting edge: molecular structure of the IL-1R-associated kinase-4 death domain and its implications for TLR signaling. *J. Immunol.* **175,** 4175–4179 (2005).
31. Liepinsh, E., Ilag, L. L., Otting, G. & Ibanez, C. F. NMR structure of the death domain of the p75 neurotrophin receptor. *EMBO J.* **16,** 4999–5005 (1997).
32. Park, H. H. & Wu, H. Crystal structure of RAIDD death domain implicates potential mechanism of PIDDosome assembly. *J. Mol. Biol.* **357,** 358–364 (2006).
33. Marti-Renom, M. A., Madhusudhan, M. S. & Sali, A. Alignment of protein sequences by their profiles. *Protein Sci.* **13,** 1071–1087 (2004).
34. Kleywegt, G. J. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr. D* **52,** 842–857 (1996).
35. DeLano, W. L. *MacPyMOL: A PyMOL-based Molecular Graphics Application for MacOS X.* (DeLano Scientific LLC, Palo Alto, California, USA, 2007).

LETTERS

Interferon modulation of cellular microRNAs as an antiviral mechanism

Irene M. Pedersen[1], Guofeng Cheng[3], Stefan Wieland[3], Stefano Volinia[4], Carlo M. Croce[4], Francis V. Chisari[3] & Michael David[1,2]

RNA interference through non-coding microRNAs (miRNAs) represents a vital component of the innate antiviral immune response in plants and invertebrate animals; however, a role for cellular miRNAs in the defence against viral infection in mammalian organisms has thus far remained elusive[1]. Here we show that interferon beta (IFNβ) rapidly modulates the expression of numerous cellular miRNAs, and that eight of these IFNβ-induced miRNAs have sequence-predicted targets within the hepatitis C virus (HCV) genomic RNA. The introduction of synthetic miRNA-mimics corresponding to these IFNβ-induced miRNAs reproduces the antiviral effects of IFNβ on HCV replication and infection, whereas neutralization of these antiviral miRNAs with anti-miRNAs reduces the antiviral effects of IFNβ against HCV. In addition, we demonstrate that IFNβ treatment leads to a significant reduction in the expression of the liver-specific miR-122, an miRNA that has been previously shown to be essential for HCV replication[2]. Therefore, our findings strongly support the notion that mammalian organisms too, through the interferon system, use cellular miRNAs to combat viral infections.

miRNAs are a class of small non-coding RNA molecules that function through post-transcriptional regulation of gene expression by a process termed RNA interference (RNAi). Over 500 miRNA-encoding genes, which seem to be exclusively transcribed by RNA polymerase II, have been identified in mammals. These primary microRNAs are processed by the enzymes Drosha/DGCR8/RNASEN into hairpin-loop-containing pre-miRNAs, which are then subject to nuclear export via exportin 5. Further enzymatic processing of the pre-miRNAs by Dicer leads to a mature miRNA duplex that is loaded into the RNA-induced silencing complex (RISC), in which the miRNA guides RISC to complementary messenger RNAs. On the basis of the degree of homology between the miRNA and the mRNA, RISC can inhibit mRNA function by either promoting its cleavage or by inhibiting its translation[3,4]. Particularly, the sequence complementarity in the 6–8 base pair 'seed region' at the 5′ end of the miRNA–mRNA heteroduplex seems to determine the specificity of miRNA–targetRNA interactions[5].

RNAi-mediated targeting of viral RNAs was first recognized in plants as an antiviral defence mechanism, but it is now apparent that invertebrates also use RNAi to combat viral infection[1,6,7]. However, thus far, no evidence has been presented in support of a protective RNAi-based antiviral response in mammalian cells. Rather, the hypothesis has been proposed that "the extremely potent interferon system has displaced RNAi as the key defence against virus infection in mammalian cells"[1]. Indeed, several type I interferon (IFNα/β)-regulated gene products such as protein kinase R, the 2'-5' OA synthase/RNase L system, the adenosine deaminase ADAR1 or the Mx GTPases are important contributors to the antiviral properties of these cytokines[8,9]. However, the possibility that IFNα/β might induce cellular miRNAs that target viral transcripts and thereby use RNAi as part of their arsenal against invading viruses has been left unexplored. To test whether IFNα/β could alter the expression of cellular miRNAs, we used microarray technology to analyse RNA derived from IFNα/β- or IFNγ-stimulated cells. This initial screening effort identified ~30 miRNAs, the expression levels of which were increased or attenuated in response to IFNα/β or IFNγ. Because we were interested in determining whether these miRNAs could potentially contribute to the antiviral effects of interferons, we performed sequence complementarity analysis of these miRNAs against viral transcripts or viral genomic RNAs with an initial focus on the crucial seed sequence. This approach revealed promising matches among several viruses, most of which harbour an RNA-based genome. Specifically, eight of the IFNβ-induced miRNAs (miR-1, miR-30, miR-128, miR-196, miR-296, miR-351, miR-431 and miR-448) displayed nearly perfect complementarity in their seed sequences with hepatitis C virus (HCV) RNA genomes. This finding was rather intriguing considering that IFNα and IFNβ are the most common treatment regimen for HCV infection[10,11]. Interestingly, a similar analysis of the hepatitis B virus (a DNA virus) yielded no significant matches.

HCV is the sole member of the hepacivirus genus of the Flaviviridiae family, and is represented by six major genotypes. The virion contains a 9.6 kb single-stranded RNA genome of positive polarity, with highly invariant 5′ and 3′ untranslated regions[12,13]. After virus entry into the host cell, the viral genome is uncoated and serves as a template for the translation of a single polyprotein, which is subsequently processed by host and viral proteases. The non-structural viral proteins then initiate the synthesis of a negative-strand RNA, which serves as a replication template for the generation of new positive-strand viral genomes[12,13]. Sequence alignment of the six HCV genotypes illustrated that the putative miRNA target sites for the IFNβ-induced miRNAs are located in areas strictly conserved among all HCV genotypes (see examples in Supplementary Fig. 1a) as well as in regions that differ between the genotypes such that high seed-sequence complementarity occurs only with selected HCV genotypes (see examples in Supplementary Fig. 1b).

To verify our microarray studies we analysed miRNA induction in response to IFNβ in the human hepatoma cell line Huh7, as well as in freshly isolated primary murine hepatocytes by quantitative PCR (qPCR). As shown in Fig. 1a, b, IFNβ treatment resulted in a similar induction of the prospective antiviral miRNAs in both cell types to that observed in ISG54, a well-characterized IFNα/β-regulated gene[14]. Two miRNAs (miR-125 and miR-142) that were found

[1]Department of Molecular Biology, Division of Biological Sciences, and [2]Moores Cancer Center, University of California San Diego, La Jolla, California 92093, USA. [3]Division of Experimental Pathology, The Scripps Research Institute, La Jolla, California 92037, USA. [4]Department of Molecular Virology, Immunology & Medical Genetics, Ohio State University, Columbus, Ohio 43210, USA.

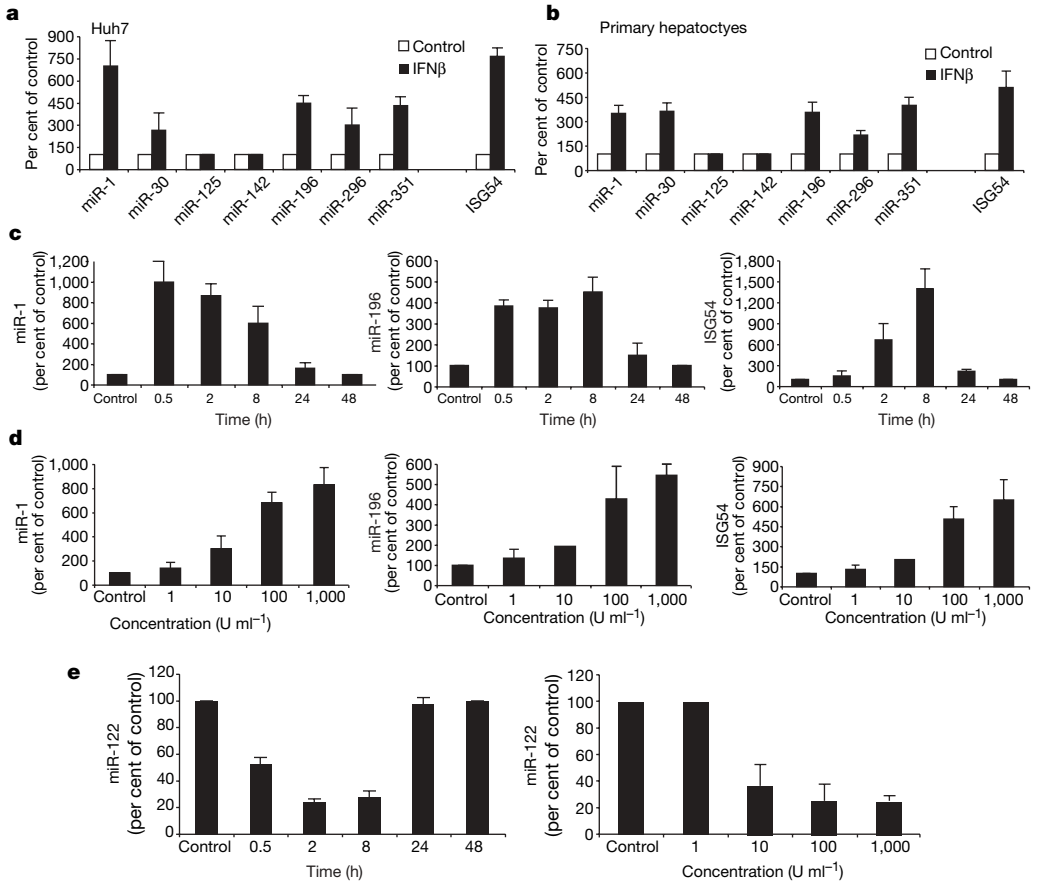


Figure 1 | Regulation of miRNA expression by IFNβ in Huh7 cells and primary hepatocytes. **a**, **b**, Huh7 cells (**a**) or primary hepatocytes (**b**) were stimulated with 100 U ml^{-1} IFNβ for 2 h, and the indicated miRNAs were quantified by qPCR. ISG54 induction is shown for comparison. **c**, Time course of miRNA induction by IFNβ: Huh7 cells were stimulated with 100 U ml^{-1} IFNβ for the indicated times, and miR-1, miR-196 or ISG54 expression was quantified by qPCR. **d**, Dose–response analysis of miRNA induction by IFNβ: Huh7 cells were stimulated with the indicated doses of IFNβ for 2 h, and miR-1, miR-196 or ISG54 expression was quantified by qPCR. **e**, Time course and dose–response analysis of miR-122 downregulation by IFNβ: Huh7 cells were stimulated as described in **c** and **d**, and miR-122 was quantified by qPCR. Error bars, means ± s.d. of at least four independent experiments.

IFNβ-unresponsive in the microarray analysis were included as negative controls. We next performed kinetic, as well as dose–response, analyses of the induction of miRNAs by IFNβ. Time course analysis revealed that induction of miR-1 and miR-196 occurs very rapidly, with peak concentrations within 30 min, and thus even precedes the upregulation of ISG54 (Fig. 1c). In a similar manner to ISG54 induction, upregulation of miR-1 and miR-196 followed a classical dose–response curve between 1 and 1,000 U ml^{-1} IFNβ (Fig. 1d), and could be blocked by actinomycin D (data not shown), indicating that the increased levels of miRNAs are the result of transcriptional induction.

miR-122 is specifically expressed in the liver, and previous studies using anti-miRNAs elegantly demonstrated that miR-122, and consequently Dicer[15], are essential for HCV replication[2]. We therefore tested whether miR-122 was also subject to regulation by IFNβ. As shown in Fig. 1e, IFNβ-stimulation of Huh7 cells resulted in a transient, but pronounced (~80%) down-modulation of miR-122. In a similar manner to miRNA induction, 100 U ml^{-1} IFNβ induced maximal attenuation of miR-122 expression, and no additional effect was observed with increased IFNβ concentrations (Fig. 1e).

To evaluate whether the eight IFNβ-induced miRNAs with sequence matches in the HCV genome are indeed capable of inhibiting HCV replication, we transfected synthetic miRNA-mimics corresponding to these miRNAs into Huh-7 cells that harbour an autonomously replicating, dicistronic full-length HCV replicon[16,17]. An anti-miRNA against miR-122 served as a positive control, because it had previously been shown to significantly reduce the abundance of replicon RNA in this system[2]. As expected, non-specific control miRNAs (individually or combined) or anti-miRNA oligos did not alter the amounts of HCV replicon RNA, whereas introduction of anti-miR-122 resulted in a ~70% reduction in viral RNA levels, as previously reported (Fig. 2a). Transfection of the eight candidate miRNAs individually revealed that miRNAs miR-196, miR-296, miR-351, miR-431 and miR-448 were indeed able to substantially

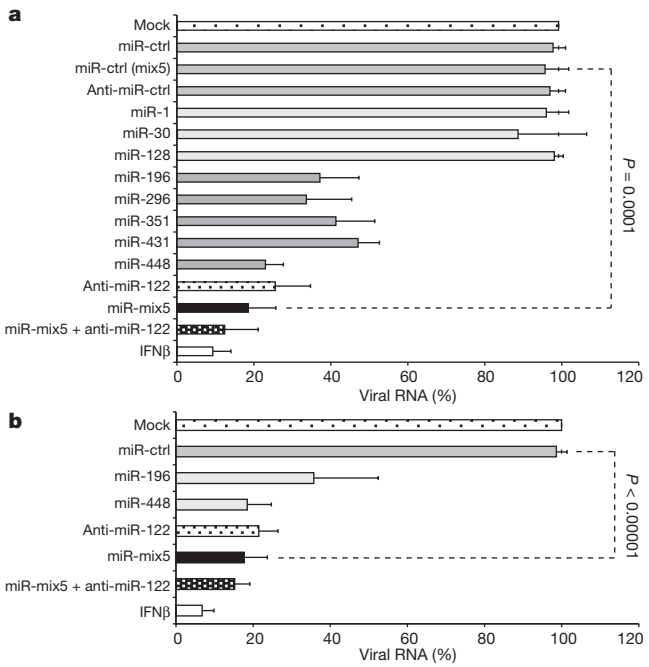
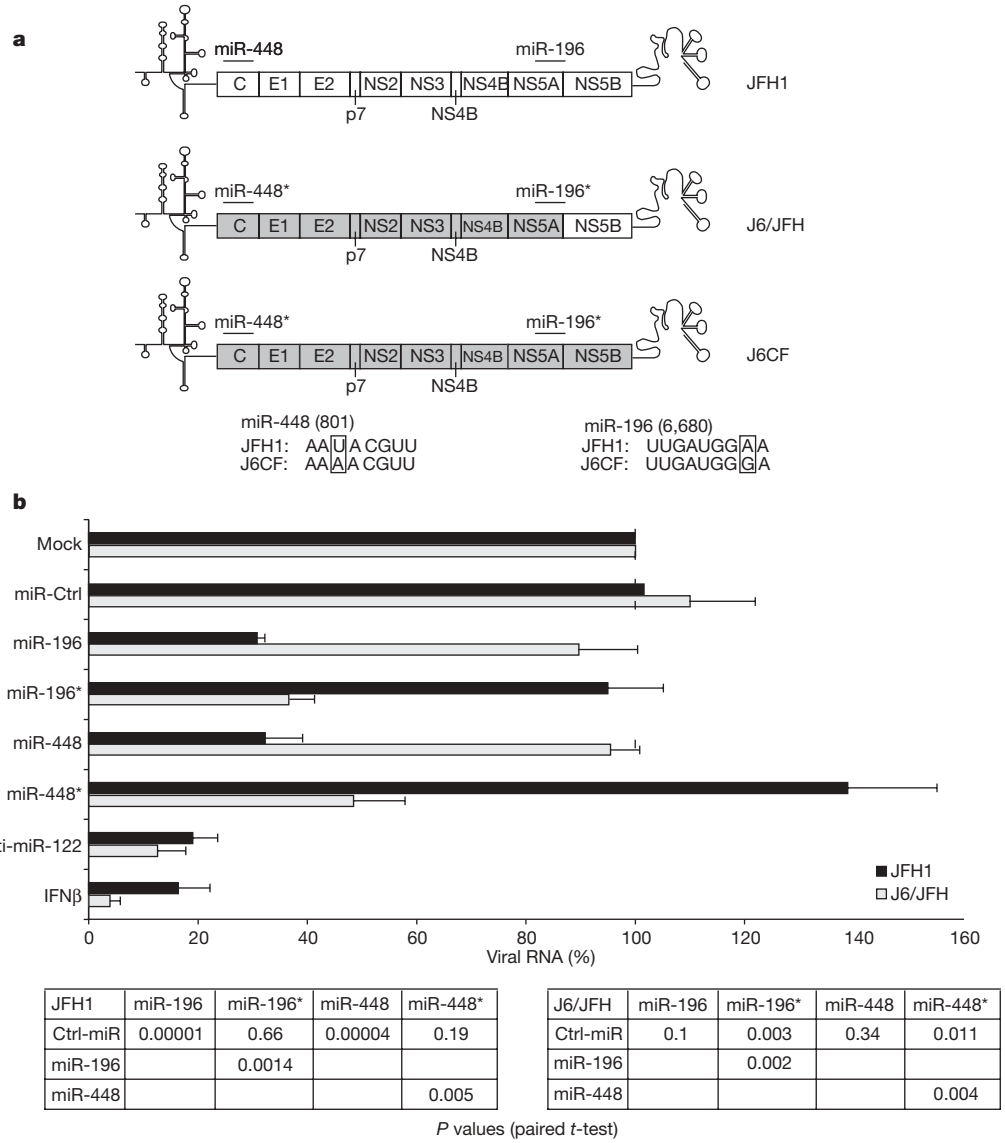


Figure 2 | IFNβ-induced miRNAs display anti-viral activity against HCV. **a**, JFH1-replicon-containing Huh7 cells were transfected with single non-specific miRNA-mimics (miR-ctrl), a pool of control miRNAs (miR-ctrl(mix5)) or anti-miRNAs (anti-miR-ctrl), or specific miRNA-mimics corresponding to the eight IFNβ-induced miRNAs, or specific anti-miRNAs. In addition, a combination of the five miRNAs that displayed anti-viral activity individually was used (miR-mix5) with or without anti-miR-122, as indicated, and HCV RNA was quantified by qPCR after 48 h. **b**, Same as **a**, except Huh7 cells were infected with live JFH-1 virus for 48 h (error bars, means ± s.d. of at least four independent experiments; *P* values are from paired Student's *t*-tests).

attenuate viral replication, whereas miRNAs miR-1, miR-30 and miR-128 were without effect (Fig. 2a). Transfection of a mixture of the five functional miRNA-mimics yielded a >80% reduction in viral RNA load. As IFNβ induces miRNAs miR-196, miR-296, miR-351, miR-431 and miR-448 but downregulates miR-122, we decided to imitate this cellular response by transfecting the mix of miRNA-mimics in combination with anti-miR-122. Indeed, a small, but reproducible additive inhibitory effect on HCV replication was observed that was similar in efficacy to treatment of the cells with IFNβ (Fig. 2a). Virtually identical results were obtained when Huh7 cells were infected with live JFH1 HCV and either viral RNA levels (Fig. 2b) or viral foci formation (Supplementary Fig. 2) were analysed, further corroborating the fact that the IFNβ-induced alterations in the miRNA expression profile endow the cells with a pronounced antiviral state.

To investigate whether these anti-viral miRNAs are targeting the HCV genome rather than inducing a non-specific antiviral state through alteration of cellular gene expression, we took advantage of the fact that the J6CF HCV molecular clone has single-nucleotide variations in the predicted target seed sequences for miR-196 and miR-448 as opposed to JFH1 (Fig. 3a). As J6CF is non-infectious *in vitro*, we employed an infection-competent J6CF/JFH1 chimaeric virus that carried the 'mutant' miR-448 and miR-196 target sites (chimaera J6/JFH) (Fig. 3a). As anticipated, miR-196 was effective against JFH1, but not against J6/JFH, which contained the 'mutant' target site. Similarly, miR-448 only inhibited the replication of JFH1 containing the 'correct' target site, but was ineffective against J6/JFH (Fig. 3b). Introduction of a compensatory single nucleotide change into miR-196 and miR-448 (designated miR-196* and miR-448*) to match their seed sequence to J6/JFH yielded a reversed efficacy profile compared to the wild-type miRNAs when they were tested against the two viruses. As such, both miR-196* and miR-448* were unable to subdue replication of JFH1, but clearly inhibited replication of J6/JFH (Fig. 3b). Together, these results suggest that at least miR-196 and miR-448 are directly targeting the HCV genomic RNA.

Although the above described experiments demonstrate that IFNβ-induced miRNAs, in conjunction with the downregulation of miR-122, are sufficient to induce an antiviral state, the question remained whether this modulation of cellular miRNA levels was necessary for IFNβ to prevent HCV replication. To this end, we decided to counteract the IFNβ-elicited changes in miRNA expression by transfecting anti-miRNAs against miRNAs miR-196, miR-296, miR-351, miR-431 and miR-448, with and without the inclusion of an miR-122 mimic. IFNβ treatment leads to a >90% reduction in the amount of viral HCV replicon RNA and this inhibition is unaffected by transfected non-specific control anti-miRNAs. As shown in Fig. 4, introduction of the anti-miRNA mix, or of the miR-122 mimic separately, attenuated the IFNβ effect to ~75% inhibition. Co-transfection of the anti-miRNA mix and the miR-122 mimic further reduced the efficacy of IFNβ to ~50%, indicating that



JFH1	miR-196	miR-196*	miR-448	miR-448*
Ctrl-miR	0.00001	0.66	0.00004	0.19
miR-196		0.0014		
miR-448				0.005

J6/JFH	miR-196	miR-196*	miR-448	miR-448*
Ctrl-miR	0.1	0.003	0.34	0.011
miR-196		0.002		
miR-448				0.004

P values (paired *t*-test)

Figure 3 | IFNβ-induced miRNAs directly target viral genomic RNA. **a**, **b**, An infectious chimaeric virus was constructed from JFH1 and J6CF as shown in **a**; the numbers in brackets indicate the location in the virion (nucleotides) that the miRNA binds. The Huh7-cell subclone Huh7.5.1c2 was transfected with either miR-196 or miR-448, or with the mutant miR-196* and miR-448* harbouring a compensatory mutation in the seed sequence as outlined in **a**. Transfected Huh7.5.1c2 cells were infected with JFH1(D183) (ref. 23) or chimaeric J6/JFH, and HCV RNA was quantified by qPCR after 24 or 60 h post infection, respectively, during the phase of exponential viral RNA amplification, to accommodate the difference in replication kinetics between the two viruses (error bars, means ± s.e.m. of 8 independent experiments; *P* values are from paired Student's *t*-tests).

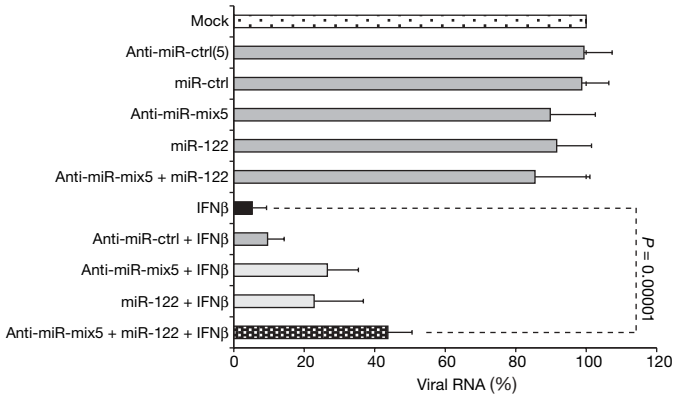


Figure 4 | IFNβ-induced miRNAs mediate anti-viral IFNβ responses against HCV. JFH1-replicon-containing Huh7 cells were transfected with non-specific miRNA-mimics (miR-ctrl), or non-specific anti-miRNAs (anti-miR-ctrl), or a pool of anti-miRNAs (anti-miR-ctrl(5)), or a combination of anti-miRNA complementary to the five IFNβ-induced miRNAs and with potent anti-viral effect (anti-miRNA-mix5), and/or with specific miRNAs and anti-miRNAs, as indicated, before stimulation with IFNβ for 48 h. HCV RNA was quantified by qPCR (means ± s.d. of at least four independent experiments; *P* values are from paired Student's *t*-tests).

modulation of the expression levels of the identified miRNAs has an important, albeit not exclusive, role in the antiviral effects of IFNβ against HCV. It remains to be determined whether alterations in the expression levels of the identified cellular miRNAs account for only a part of the antiviral response, or if additional unidentified miRNAs that have not been 'neutralized' by anti-miRNAs mediate the remaining antiviral IFNβ effects.

In summary, our results demonstrate that IFNα/β upregulates several cellular miRNAs that are capable of inhibiting HCV replication and infection. In addition, downregulation of miR-122 in response to IFNβ further contributes to the antiviral effects of this cytokine. These findings not only offer a new model of the host defence mechanisms that exist in mammalian cells, but also add a new component to the antiviral arsenal employed by interferons. Furthermore, although ample documentation is present in the literature for the developmentally regulated or tissue-specific expression of miRNAs and their dysregulation in malignant cells[18,19], little evidence existed for a direct and immediate transcriptional regulation of miRNAs by an endogenous ligand such as a cytokine or growth factor. Our results provide some of the first evidence for rapid changes in cellular miRNA levels in response to ligand stimulation. During the preparation of this manuscript, it was reported that IFNβ induced upregulation of miR-155 (ref. 20), a miRNA with, as of yet, no identified targets. The delayed induction kinetics observed in their studies supports the model that miR-155 upregulation is mediated by TNFα as part of an autocrine response elicited by IFNβ, rather than by IFNβ itself[20]. Nevertheless, this study and our present report provide the first examples of rapid modulation of cellular miRNAs as components of the mammalian innate immune response.

METHODS SUMMARY

Cell culture. The human hepatoma-derived cell lines Huh7 and their JFH-1-replicon-containing subclone were maintained in DMEM with 10% FCS/10mM Hepes, penicillin/streptomycin and 2 mM L-glutamine. The JFH-1 full-length genomic replicon construct pFGR-JFH-1 was obtained from T. Wakita[21]. A Huh7 cell clone that stably replicates the full-length genomic HCV RNA was selected and used in these experiments, as described previously[22]. The Huh7.5.1c2 cell line was derived from curing replicon-containing Huh-7.5.1 cells by interferon treatment. Primary murine hepatocytes (BalbC) were collected after collagenase perfusion of the livers for 30 min at 37 °C.

miRNA array analysis. Microarray analysis of total RNA of interferon-stimulated lymphocytes was performed at the Ohio State University Comprehensive Cancer Center Microarray Shared Resource, as described[23].

RNA extraction. Total RNA was isolated using TriZol according to the manufacturer's instructions.

Real-time PCR. Quantification of HCV genomic RNA was performed using HCV- and β-actin- or GAPDH-specific primers as described[16]. Real-time PCR-based quantification of miRNAs was performed using miRNA analysis kits specific for each individual miRNA (Applied Biosystems), according to the manufacturers instructions.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 July; accepted 29 August 2007.

1. Cullen, B. R. Is RNA interference involved in intrinsic antiviral immunity in mammals? *Nature Immunol.* **7,** 563–567 (2006).
2. Jopling, C. L., Yi, M., Lancaster, A. M., Lemon, S. M. & Sarnow, P. Modulation of hepatitis C virus RNA abundance by a liver-specific microRNA. *Science* **309,** 1577–1581 (2005).
3. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–297 (2004).
4. Ambros, V. The functions of animal microRNAs. *Nature* **431,** 350–355 (2004).
5. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
6. Zamore, P. D. Plant RNAi: How a viral silencing suppressor inactivates siRNA. *Curr. Biol.* **14,** R198–R200 (2004).
7. Baulcombe, D. RNA silencing in plants. *Nature* **431,** 356–363 (2004).
8. Katze, M. G., He, Y. & Gale, M. Jr. Viruses and interferon: a fight for supremacy. *Nature Rev. Immunol.* **2,** 675–687 (2002).
9. Samuel, C. E. Antiviral actions of interferons. *Clin. Microbiol. Rev.* **14,** 778–809 (2001).
10. Liang, T. J., Rehermann, B., Seeff, L. B. & Hoofnagle, J. H. Pathogenesis, natural history, treatment, and prevention of hepatitis C. *Ann. Intern. Med.* **132,** 296–305 (2000).
11. Lauer, G. M. & Walker, B. D. Hepatitis C virus infection. *N. Engl. J. Med.* **345,** 41–52 (2001).
12. Wieland, S. F. & Chisari, F. V. Stealth and cunning: hepatitis B and hepatitis C viruses. *J. Virol.* **79,** 9369–9380 (2005).
13. Rehermann, B. & Nascimbeni, M. Immunology of hepatitis B virus and hepatitis C virus infection. *Nature Rev. Immunol.* **5,** 215–229 (2005).
14. Larner, A. C. *et al.* Transcriptional induction of two genes in human cells by beta interferon. *Proc. Natl Acad. Sci. USA* **81,** 6733–6737 (1984).
15. Randall, G. *et al.* Cellular cofactors affecting hepatitis C virus infection and replication. *Proc. Natl Acad. Sci. USA* **104,** 12884–12889 (2007).
16. Zhong, J. *et al.* Robust hepatitis C virus infection *in vitro*. *Proc. Natl Acad. Sci. USA* **102,** 9294–9299 (2005).
17. Moradpour, D. *et al.* Insertion of green fluorescent protein into nonstructural protein 5A allows direct visualization of functional hepatitis C virus replication complexes. *J. Virol.* **78,** 7400–7409 (2004).
18. Calin, G. A. & Croce, C. M. MicroRNA–cancer connection: the beginning of a new tale. *Cancer Res.* **66,** 7390–7394 (2006).
19. Calin, G. A. & Croce, C. M. MicroRNA signatures in human cancers. *Nature Rev. Cancer* **6,** 857–866 (2006).
20. O'Connell, R. M., Taganov, K. D., Boldin, M. P., Cheng, G. & Baltimore, D. MicroRNA-155 is induced during the macrophage inflammatory response. *Proc. Natl Acad. Sci. USA* **104,** 1604–1609 (2007).
21. Kato, T. *et al.* Efficient replication of the genotype 2a hepatitis C virus subgenomic replicon. *Gastroenterology* **125,** 1808–1817 (2003).
22. Cheng, G., Zhong, J. & Chisari, F. V. Inhibition of dsRNA-induced signaling in hepatitis C virus-infected cells by NS3 protease-dependent and -independent mechanisms. *Proc. Natl Acad. Sci. USA* **103,** 8499–8504 (2006).
23. Liu, C. G. *et al.* An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc. Natl Acad. Sci. USA* **101,** 9740–9744 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Calin (OSU) for help with the microarray studies, D. Burton (TSRI) for E2-antibodies, and J. Zhong (TSRI) for many discussions. This work was supported by the NIH (I.M.P.), a private donation from C. Evans to F.V.C, and the NIH and UCSD Academic Senate Funding (M.D.).

Author Contributions G.C. and S.W. contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.D. (midavid@ucsd.edu).

METHODS

Transfection of miRNA mimics and anti-miRNAs into Huh7 or Huh7 replicon cells. When JFH1-replicon-containing Huh7 cells (70% confluent) were used for transfection experiments, constant total amounts of miRNA mimics or anti-miRNAs (Dharmacon and Applied Biosystems, respectively) were combined with Mirus transfection reagent (Mirus Bio/Fisher Scientific) according to the manufacturers instructions and added to the cells (transfection efficiency >90%). Where indicated, IFNβ (DBL Biomedical Laboratories) was added to the cultures after 8 h, and cells were harvested 48 h post transfection. Viral replication was determined by real-time PCR analysis of viral genomic RNA. Similarly, for experiments using infectious HCV, Huh7 cells were transfected with miRNAs and anti-miRNAs, as described above. HCV infection (multiplicity of infection = 0.05) was performed 8 h post transfection, and IFNβ was added where indicated to the cultures at the same time. Cells were harvested at the indicated time points during the phase of exponential viral RNA amplification, and viral replication was determined by real-time PCR analysis of viral genomic RNA, or by determination of the number and size of viral foci after 72 h.

J6CF–JFH1 chimaeric HCV genomes. Recombinant PCR was used to replace the J6CF NS5B-3′ UTR region in pCV-J6CF[24] with the corresponding sequences from JFH-1 present in pUC-vJFH[25]. First, using primers XbaJFH (GATTACGCCAAGCTTGCATGCCTGCAG), and NS5Bup (CTCCATGTCATACTCCTGGACCGGGGCTC) and primers NS5Blo (GAGCCCCGGTCCAGGAGTATGACATGGAG) and XhoJ6CF (AGGTTCCATCTCTTCCATGCCCCCCCTCG) the NS5B-3′ UTR region of JFH-1 and an NS5A-NS5B fragment containing a unique *Xho*I restriction site from J6CF were amplified, respectively. The two PCR products were mutually extended and amplified using the primers XhoJ6CF and XbaJFH, and the resultant PCR product was cloned into pGEM-Teasy (Invitrogen) yielding pTe-J6/JFH, and the insert was verified by DNA sequencing. Finally, the *Xho*I to *Xba*I fragment in pCV-J6CF was replaced by the 2.1-kb *Xho*I/*Xba*I fragment excised from pTe-J6/JFH, yielding pCV-J6CF/JFHNS5B3'UTR containing the J6/JFH chimaeric HCV genome with the NS5B-3′ UTR of J6CF replaced by the corresponding sequences from JFH-1. A similar recombinant PCR approach was used to replicate the J6CF–JFH-1 chimaeric HCV genome Jc1[26] by replacing the corresponding JFH-1 core-NS2 region in pUC-vJFH with the corresponding sequences from J6CF to yield the plasmid pUC-vJc1. Infectious JFH-1 and chimaeric viruses were produced by transfection of *in vitro* synthesized genomic HCV RNA into Huh-7.5.1 cells and virus stocks containing 10^4–10^5 focus forming units per ml (f.f.u. ml^{-1}) were prepared as described previously[27].

24. Yanagi, M., Purcell, R. H., Emerson, S. U. & Bukh, J. Hepatitis C virus: an infectious molecular clone of a second major genotype (2a) and lack of viability of intertypic 1a and 2a chimeras. *Virology* **262**, 250–263 (1999).
25. Wakita, T. *et al.* Production of infectious hepatitis C virus in tissue culture from a cloned viral genome. *Nature Med.* **11**, 791–796 (2005).
26. Pietschmann, T. *et al.* Construction and characterization of infectious intragenotypic and intergenotypic hepatitis C virus chimeras. *Proc. Natl Acad. Sci. USA* **103**, 7408–7413 (2006).
27. Zhong, J. *et al.* Robust hepatitis C virus infection *in vitro*. *Proc. Natl Acad. Sci. USA* **102**, 9294–9299 (2005).

LETTERS

Architectural and mechanistic insights into an EHD ATPase involved in membrane remodelling

Oliver Daumke[1]*†, Richard Lundmark[1]*†, Yvonne Vallis[1], Sascha Martens[1], P. Jonathan G. Butler[1] & Harvey T. McMahon[1]

The ability to actively remodel membranes in response to nucleotide hydrolysis has largely been attributed to GTPases of the dynamin superfamily, and these have been extensively studied[1]. Eps15 homology (EH)-domain-containing proteins (EHDs/RME-1/pincher) comprise a less-well-characterized class of highly conserved eukaryotic ATPases implicated in clathrin-independent endocytosis[2], and recycling from endosomes[3,4]. Here we show that EHDs share many common features with the dynamin superfamily, such as a low affinity for nucleotides, the ability to tubulate liposomes *in vitro*, oligomerization around lipid tubules in ring-like structures and stimulated nucleotide hydrolysis in response to lipid binding. We present the structure of EHD2, bound to a non-hydrolysable ATP analogue, and provide evidence consistent with a role for EHDs in nucleotide-dependent membrane remodelling *in vivo*. The nucleotide-binding domain is involved in dimerization, which creates a highly curved membrane-binding region in the dimer. Oligomerization of dimers occurs on another interface of the nucleotide-binding domain, and this allows us to model the EHD oligomer. We discuss the functional implications of the EHD2 structure for understanding membrane deformation.

EHDs comprise a highly conserved eukaryotic protein family with four members (EHD1–4) in mammals and a single member in *Caenorhabditis elegans*, *Drosophila melanogaster* and in many eukaryotic parasites, such as the genera *Plasmodium, Leishmania* and *Entamoeba*. The proteins have a molecular mass of approximately 60 kDa and contain an amino-terminal G-domain, followed by a helical domain and a carboxy-terminal EH-domain (Fig. 1a; in plant homologues the EH-domain is N-terminal). The EH-domain is known to interact with Asn-Pro-Phe (NPF) motifs in proteins involved in endocytosis. Overexpressed EHDs can be found on tubules inside cells and EHD family members (including RME-1/EHD1 and pincher) have been shown to regulate exit from the endocytic recycling compartment, TrkA-receptor-mediated macropinocytosis and other trafficking pathways[2–12]. Here we explore the structure and function of mouse EHD2 as a model for the EHD family.

Mouse full-length EHD2 was expressed in bacteria and purified to homogeneity (Supplementary Fig. 1). The purified protein was nucleotide-free, as judged by HPLC analysis, and was dimeric in analytical velocity centrifugation (Supplementary Fig. 2). It was previously reported that EHDs—despite having a predicted G-domain—bind to adenine nucleotides[13]. We confirmed these results by using isothermal titration calorimetry: the affinities for ATP-γ-S and ADP were 13 μM and 50 μM, respectively (Fig. 1b). Nucleotide binding was Mg^{2+}-dependent (data not shown). A mutation in the GKT motif in the phosphate-binding (P-)loop, T72A, prevented binding to ATP-γ-S in a manner similar to that of equivalent mutations in other GTPases. No binding signal was observed for GTP-γ-S (Fig. 1b). We still refer to the N-terminal domain as a G-domain because of sequence and fold similarity (see below) to other G-domains. We investigated membrane-binding properties of EHD2 and found efficient binding to liposomes of brain-derived lipids (Folch extract) and to 100% anionic phosphatidyl-serine (PtdSer) liposomes (Fig. 1c). However, using synthetic liposomes containing different phosphatidyl inositols (PtdIns) and more stringent conditions (only 10% PtdSer, Fig. 1c), we observed preferential binding to liposomes containing PtdIns(4,5)bisphosphate, PtdIns(4,5)P_2.

The consequence of membrane binding was analysed by electron microscopy, and we found that EHD2 deforms PtdSer liposomes in a nucleotide-independent manner into 20-nm diameter tubules and oligomerizes in ring-like structures around these tubules (Fig. 1d, and Supplementary Fig. 3). Nucleotide independence of liposome tubulation *in vitro* was also observed for dynamin[1]. No noticeable tubule fission or alteration in tubule diameter was found for EHD2 in the presence of ATP. Frequently, we observed a complex network of connected tubules that had an extensive surface area, implying that there is considerable fusion occurring between liposomes (Supplementary Fig. 3). We also saw a few instances where EHD2 oligomeric rings were of variable diameter (Fig. 1d, right panel) suggesting that the interface used for EHD2 oligomerization is rather flexible. Folch liposomes were also tubulated by EHD2 in a nucleotide-independent manner (data not shown). However, with synthetic liposomes containing only 2.5% PtdIns(4,5)P_2 and in the absence of PtdSer, the amount of EHD binding was reduced, and in this less-favourable binding condition we only observed tubulation in the presence of ATP-γ-S and ADP, but not in the absence of nucleotides (Supplementary Fig. 3).

When enhanced green fluorescent protein (EGFP)-tagged EHD2 was overexpressed in HeLa cells it marked punctate and tubular structures (Fig. 1e, endogenous EHD2 in various cell lines shows a similar peripheral distribution[5,14]). By total internal reflection fluorescence microscopy, those structures were mainly found close to the plasma membrane (data not shown), consistent with the observed PtdIns(4,5)P_2 specificity. Although the nucleotide-free T72A mutant bound to Folch liposomes *in vitro* (Fig. 1c), it showed a cytoplasmic distribution when overexpressed *in vivo* (Fig. 1e), indicating that nucleotide binding is required for oligomerization *in vivo*, in agreement with previous results[3].

The effect of membrane binding on the ATPase activity of EHD2 was monitored under multiple-turnover conditions (tenfold excess of ATP over EHD2), in the presence and absence of Folch liposomes (Fig. 1f, and Supplementary Fig. 4). The intrinsic/background

[1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 0QH, UK. †Present addresses: Max-Delbrück-Centrum für Molekulare Medizin (MDC), Robert-Rössle-Str. 10, 13092 Berlin, Germany (O.D.); Umeå University, Department of Medical Biochemistry and Biophysics, 90187 Umeå, Sweden (R.L.).
*These authors contributed equally to this work.

ATPase activity is extremely slow ($k_{obs} = 0.7\,h^{-1}$) but is stimulated by liposome binding and can be maximally activated eightfold in the presence of saturating Folch or PtdSer liposomes ($k_{obs} = 5.6\,h^{-1}$). In contrast to GBP1 (ref. 15), we did not observe hydrolysis to nucleoside mono-phosphate. GTP was not hydrolysed in the presence or absence of Folch liposomes, and the T72A mutant did not show membrane-stimulated ATPase activity. EHD2 displays a 600-fold slower stimulated nucleotide hydrolysis than dynamin, which hydrolyses GTP under our standard conditions (see Methods) with a k_{cat} of $\approx 1\,s^{-1}$.

To obtain structural insights, we solved the crystal structure of EHD2 in the presence of the non-hydrolysable ATP analogue AMP-PNP (see Methods, Table 1 and Supplementary Table 1 for the statistics). The nucleotide-binding domain of EHD2 possesses a typical G-domain fold with a central β-sheet surrounded by α-helices (Fig. 2a, b, and Supplementary Fig. 5). An AMP-PNP molecule occupies the canonical nucleotide-binding site. In comparison to the Ras-like G-domain, EHD2 contains an insertion of two additional β-strands in the switch I region, which are also present in the G-domain of dynamin[16] (Fig. 2b). Surprisingly, switch I and switch II are not well ordered in this dimeric protein (Fig. 2a). Residues 110–130, which are distal to switch I, are disordered and contain predicted EH-domain binding motifs, KPFxxxNPF. In agreement with our biochemical analysis, EHD2 crystallizes as a dimer, and the dimer axis corresponds to a crystallographic two-fold axis (Fig. 2a). Dimerization is mediated by a highly conserved, mostly hydrophobic interface of approximately 2,100 Å^2 in the G-domain (Supplementary Fig. 6). At the centre of the interface, the entirely conserved W238 in helix α6 is buried in a hydrophobic pocket, and mutations of this residue render the protein insoluble (data not shown). This novel interface is highly conserved among EHD family members and involves a different face of the G-domain to the dimer interfaces from the structurally characterized GBP1 and bacterial dynamin-like protein (BDLP) dimers[15,17].

Table 1 | X-ray refinement statistics

Resolution (Å)	20–3.1
Number of reflections	12,091
R_{work}/R_{free} (%)	23.6/27.6
Number of atoms	
Protein	3,765
Ligand/ion	33
Water	6
B-factors	
Protein	61 Å^2
Ligand/ion	50 Å^2
Water	51.5 Å^2
Root mean squared deviations	
Bond lengths (Å)	0.019
Bond angles (°)	1.119

The helical domain is composed of helix α1 and α2 from the N-terminal region (residues 18–55, which follow disordered residues 1–18) and helices α8 to α12 (residues 285–400) following the G-domain (Fig. 2a,b). Helix α8 is the organizing scaffold against which most of the other helices fold. It also has extensive contacts with the G-domain. The dimeric G-domain, together with the helical region, adopts a scissor shape, in which the membrane is proposed to bind between the blades (see later). Following the middle domain, a 40-residue linker connects the helical domain with the C-terminal EH-domain (residues 443–543). The EH-domain of EHD2 is similar to the previously determined second EH-domain of Eps15 solved by NMR studies[18,19], with a root mean squared deviation of 1.5 Å for the main-chain atoms (Fig. 2c). It is built of two closely packed perpendicular EF hands, which are connected by a short β-sheet. We included a Ca^{2+}-ion in the second EF hand, which is ligated by four oxygens from acidic side chains and one main-chain carbonyl oxygen

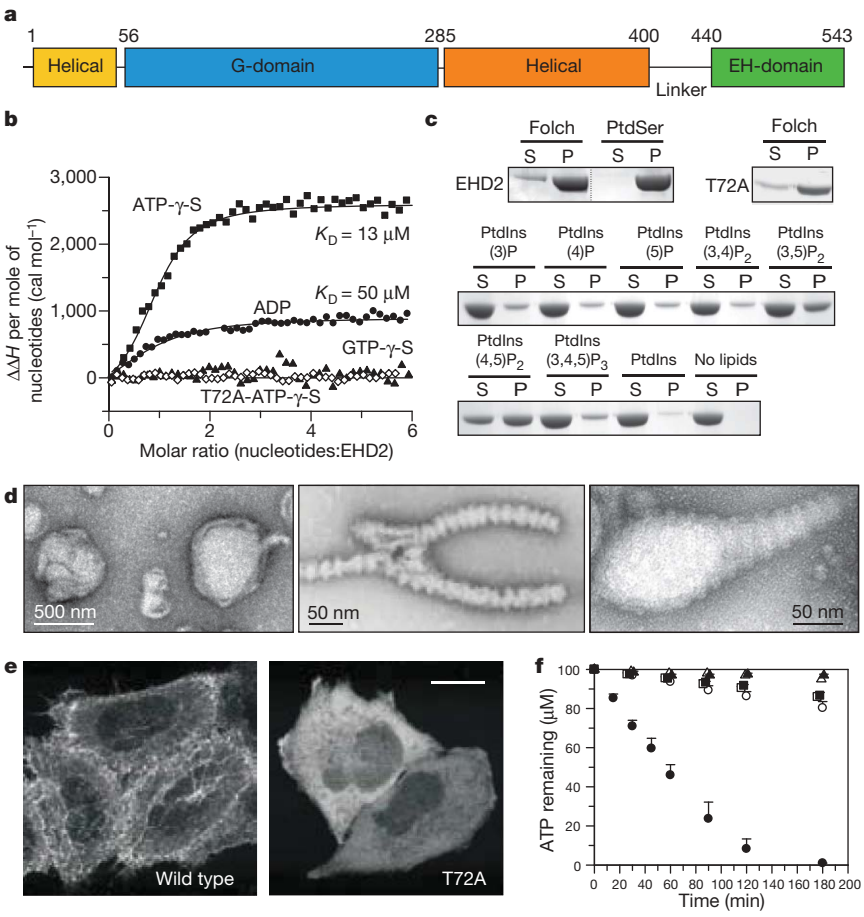


Figure 1 | EHD2 shares common properties with the dynamin superfamily. **a**, Domain structure of EHD proteins (numbering from mouse EHD2 amino acids). **b**, EHD2 binds to adenine nucleotides, as determined by isothermal titration calorimetry. For clarity, $\Delta\Delta H = \Delta H_n - \Delta H_1$ is plotted. **c**, Coomassie-stained gels of liposome co-sedimentation assays in the presence of 1 mM ATP-γ-S using 0.8-μm-filtered Folch, 100% PtdSer or synthetic liposomes containing 10% of the indicated PtdIns (plus 70% Ptd choline, 10% PtdSer, 10% cholesterol). S, supernatant; P, pelleted fraction. **d**, Electron micrographs of negatively stained PtdSer liposomes in the absence (left panel) or presence (middle, right panels) of EHD2 and 1 mM ATP-γ-S. The right panel shows an intermediate in the tubulation process, surrounded by EHD2 rings of variable diameter. **e**, Amino-terminally EGFP-tagged EHD2 wild type and the T72A mutant were overexpressed in HeLa cells. Confocal images were acquired close to the basal cell surface; the scale bar is 10 μm. **f**, Nucleotide hydrolysis by EHD2 was measured by HPLC. Intrinsic reactions in absence of lipids are open symbols (mean ± s.d., $n = 2$) and stimulated reactions are filled symbols (mean ± s.d., $n = 3$). Whereas the intrinsic ATP reaction was eightfold-stimulated by Folch lipids (open versus filled circles), GTP hydrolysis was not stimulated (open versus filled triangles). The nucleotide-free T72A mutant did not show stimulation of ATP hydrolysis (open versus filled squares).

(Supplementary Fig. 7a). The EH-domain is localized on top of the G-domain (the top-site position) with a buried interface of 1,600 Å^2. Eighteen disordered residues connecting the EH-domain to the helical domain mean that it is ambiguous which EH-domain connects to which helical domain. We assign the EH-domains to opposing monomers (red Elt-domain belongs to the red helical domain in Fig. 2a) because the last visible residue of the linker from the helical domain is closer to the opposite EH-domain (distance 29 Å) than to the superjacent EH-domain (distance of 34 Å), and, in this latter case, the linker would have to wind around the EH-domain and thus one would expect to see some parts of this in the structure (Fig. 2a). Unexpectedly, the peptide-binding sites of both EH-domains are occupied by a GPF motif (residues 420–422) from the linker region (Fig. 2a and c). The GPF motif adopts a similar conformation to an NPF-containing peptide bound to the EH2 domain of Eps15 (ref. 18), involving a tight turn with F422 projecting into a hydrophobic pocket that is lined by W490 at its base (Fig. 2c and Supplementary Fig. 7a). R536 in the C-terminal tail of the EH-domain points into the active site of the G-domain below and forms a hydrogen bond to D222 from the NKxD motif, which in other GTPases is responsible for specific recognition of the guanosine base (see Supplementary Fig. 8 for details). Furthermore, M223 sterically prevents binding of GTP, thus explaining the ATP specificity of EHD2.

To test if a polybasic stretch close to the tip of the helical domain was involved in lipid binding, we monitored lipid binding of K327D, K328D and also F322A (Fig. 3a). These mutants showed reduced liposome binding and ATPase activity in the presence of Folch liposomes (Fig. 3b, c). *In vivo*, F322A, and every lysine to aspartate mutation of K324, K327, K328, K329, led to a cytoplasmic distribution of the protein (Fig. 3d, and data not shown). Thus, the lipid interaction site of EHD2 is composed of two closely apposing lipid-binding sites in the dimer, leading to a highly curved membrane-binding interface (Fig. 3a, and Supplementary Fig. 9). Under stringent binding conditions, EHD2 indeed showed a binding preference for very small liposomes, consistent with this curvature, but only in the nucleotide-free form (Supplementary Fig. 9). In the nucleotide-bound form, EHD binding was not curvature-sensitive, and this is probably due to oligomer formation along an axis perpendicular to the high curvature (see below).

In EHD2, the phosphate groups of the AMP-PNP molecule are occluded from the exterior by a 'phosphate cap' that is formed by switch I and the P-loop, which would not allow the insertion of a catalytic residue *in trans* into the catalytic site (Supplementary Fig. 10). We searched for *in cis* catalytic residues and found that the switch I mutant T94A bound to ATP-γ-S with nearly wild-type affinity and oligomerized around PtdSer liposomes, but did not show any membrane-stimulated ATPase activity (Fig. 3e). This is consistent with a catalytic function for T94. When overexpressed in HeLa cells, the T94A mutant labelled extensive tubular structures with essentially no punctate staining (compare wild type and T94A in Fig. 3d, f and g, and Supplementary Movies 1–4), suggesting that ATP hydrolysis is involved in the break-down of tubular structures *in vivo*. We previously observed severely inhibited GTP hydrolysis and an extensive tubulation phenotype (like that found with EHD2 T94A) with dynamin 1 when T65 in switch I was mutated to alanine[20]. We analysed another mutant in switch II, in which a glutamine was introduced close to the catalytic site. I157Q hydrolysed ATP faster than wild type, even in the absence of membranes and tubulated liposomes (Fig. 3e). When overexpressed in HeLa cells this protein labelled only very short tubules and puncta (Fig. 3f, g). Many of these puncta were highly mobile as determined by live-cell microscopy (Supplementary Movies 5, 6). Thus, there is a clear correlation between the measured ATP hydrolysis rates *in vitro* and the extent of tubule formation *in vivo*. The ATPase rates might be further stimulated *in vivo*, by binding partners or modifications. Altogether, these results are consistent with a role of EHD2 ATP hydrolysis in membrane remodelling and/or the scission of membranes *in vivo*.

A highly conserved surface patch encompasses switch I, switch II and the surrounding area of the EHD2 dimer (Fig. 4a). GBP1 and BDLP use this same interface, with the same relative orientation of the G-domains, for dimerization[15,17]. Thus, the EHD2 dimer may further oligomerize into the observed rings (Fig. 1d) using this second G-domain interface. This oligomerization probably leads to the ordering of residues in the switch regions, leading to increased nucleotide hydrolysis. Four single mutations of conserved surface-exposed residues in this interface did not affect lipid binding or oligomerization on liposomes but greatly reduced the liposome-stimulated ATPase reaction (Fig. 4b, Supplementary Fig. 11a). The most drastic mutant, K193D, was further tested *in vivo* and showed extensive tubulation, whereas the E91Q mutant, which had the mildest effect on the stimulated ATPase reaction, was indistinguishable from wild type *in vivo* (Supplementary Fig. 11b). These results are

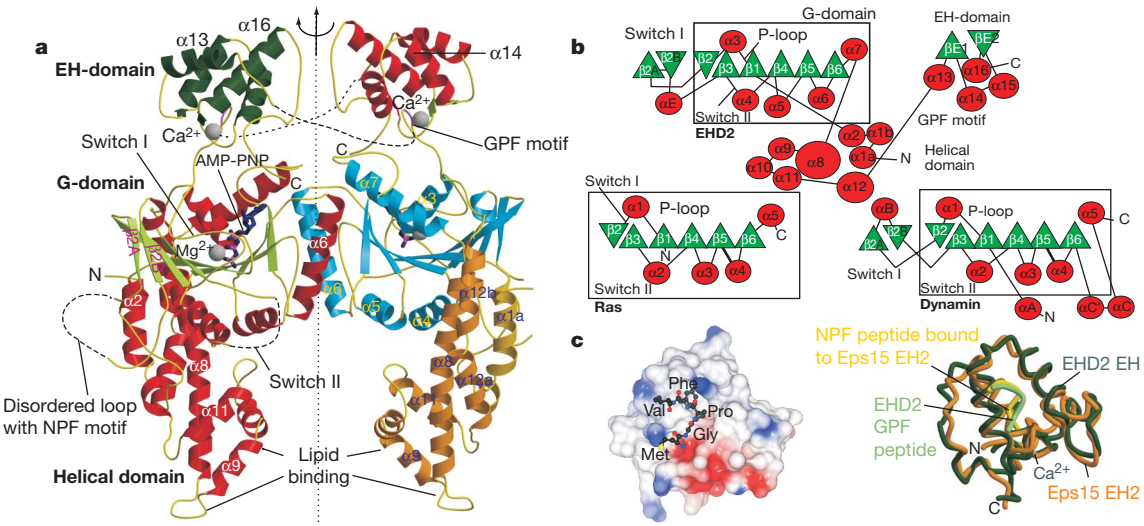
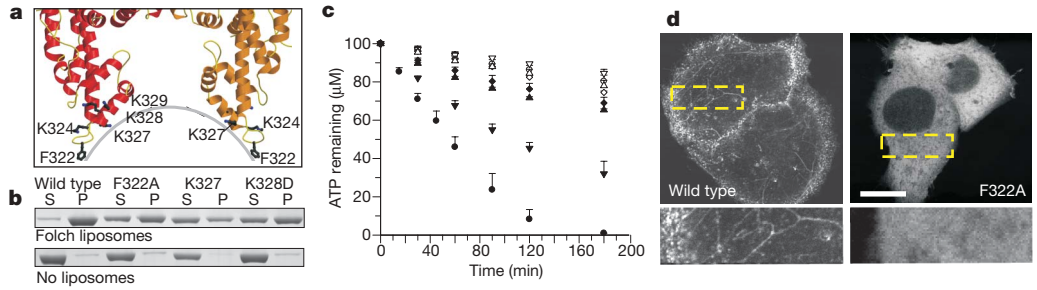


Figure 2 | The structure of EHD2. a, Ribbon-type presentation of the EHD2 dimer (PDB code 2QPT). One molecule is coloured according to the secondary structure (helices in red, β-strands in green) and the other according to the domain structure (see Fig. 1a). GPF and NPF motifs are indicated. **b**, Topology plot of EHD2 (circles represent α-helices, triangles represent β-strands). EHD2 has a dynamin-related switch I extension when compared with the classical Ras-like G-domain (boxed) and the dynamin G-domain. **c**, Left panel, electrostatic surface representation of the EHD2 EH-domain (at neutral pH: red, negative; blue, positive). F422 is penetrating in the non-charged peptide-binding pocket. Right panel, close superposition of the EHD2 EH-domain (dark-green) with the second EH-domain of Eps15 (orange) bound to an NPF-containing peptide (PDB code 1FF1)[19].



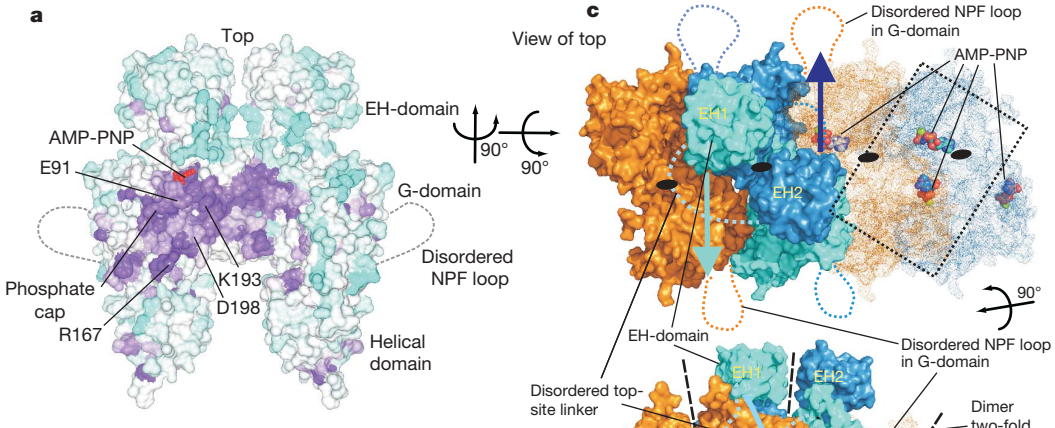
e

	K_D for ATP-γ-S, stoichiometry	Intrinsic rate (h^{-1})	Stimulated rate (h^{-1})	Tubulation *in vitro*
Wild type	13 μM, $n = 0.96$	0.66 ± 0.02	5.4 ± 0.2	+
T94A	21 μM, $n = 0.93$	0.9 ± 0.1	0.64 ± 0.03	+
I157Q	Not measured	8.8 ± 0.8	8.8 ± 0.7	+

g

	Cells expressing a low and medium level of protein (per cent of total)			Cells expressing a high level of protein (per cent of total)		
	Only puncta	Puncta and tubules	Only tubules	Only puncta	Puncta and tubules	Only tubules
Wild type	14	86	0	88	12	0
T94A	1	0	99	0	0	100
I157Q	100	0	0	100	0	0

Figure 3 | Membrane binding and the role of ATP hydrolysis. **a**, Putative membrane-binding site with residues tested for membrane binding in ball-and-stick representation. The highly curved membrane interaction site of EHD2 is indicated. **b**, Sedimentation assays in the presence (upper panel) and absence (lower panel) of Folch liposomes using wild-type EHD2 and mutants, as in Fig. 1c. **c**, Nucleotide hydrolysis of lipid binding mutants as described in Fig. 1f. The wild-type protein had a k_{obs} of 1.6 h^{-1} (open circles, intrinsic; filled circles, lipid stimulated). The F322A mutant (open versus filled inverted triangles) showed a 40% decrease in the stimulated ATPase reaction ($k_{obs} = 3.0\ h^{-1}$), and the K328D mutant (open versus filled triangles) showed a 75% reduced rate ($k_{obs} = 1.6\ h^{-1}$), whereas for the K327D mutant (open versus filled diamonds) stimulation was barely visible. **d**, EGFP-tagged F322A mutant showed a completely cytoplasmic distribution when overexpressed in HeLa cells. Scale bar, 10 μm. **e**, Affinity to ATP-γ-S was measured as in Fig. 1b. Nucleotide hydrolysis was measured as described in Fig. 1f (values represent mean ± s.e.m.; $n = 2$, intrinsic; $n = 3$, stimulated reactions). *In vitro* tubulation activity of PtdSer liposomes was analysed as described in Fig. 1d. **f**, Confocal images of HeLa cells overexpressing the indicated mutants. Scale bar, 10 μm. **g**, Quantification of the overexpression phenotypes from Fig. 3f. For each construct, three independent experiments with ≈50 cells per experiment were analysed.



b

	Intrinsic rate (h^{-1})	Stimulated rate (h^{-1})	Fold stimulation	Tubulation *in vitro*
Wild type	0.66 ± 0.02	5.4 ± 0.2	8.0	+
E91Q	0.50 ± 0.04	1.7 ± 0.1	3.4	+
R167E	1.8 ± 0.1	2.0 ± 0.2	1.1	+
K193D	0.45 ± 0.02	0.25 ± 0.03	–	+
D198R	0.53 ± 0.02	1.0 ± 0.1	1.9	+
EHD2 ΔEH	0.24 ± 0.03	0.13 ± 0.03	–	–
EHD2 F122A/F128A	0.71 ± 0.03	0.72 ± 0.03	1	–

Figure 4 | The EHD2 oligomer. **a**, Surface conservation plot of the EHD2 dimer (in the same orientation as in Fig. 2a, using the alignment in Supplementary Fig. 5) with conserved residues in purple and non-conserved residues in cyan. Conserved surface-exposed residues mutated in the following experiments are indicated. The relative orientation of the dimers in Fig. 4c is also indicated by arrows. **b**, Nucleotide hydrolysis (as in Fig. 1f) for mutants in the proposed oligomerization interface. Note that mutant R167E (in switch II) showed an increased basal ATPase activity, which was not further stimulated by Folch liposomes. Tubulation of PtdSer liposomes was assayed by negative-stain electron microscopy as in Fig. 1d. **c**, Model of the EHD2 oligomer. Upper panel, top view of the proposed oligomer, limited to four EHD2 dimers, in which the two-fold axis of the dimers is indicated by the black oval. Only the blue–cyan dimer has its EH-domains included so that the proposed movement of the EH-domains from the top-site linker to the side-site of the G-domain can be better visualized (see arrows for movements). The AMP-PNP molecules face each other in a head-to-head fashion diagonally across each dimer interface (see dotted box), as seen in the GBP1 and BDLP dimers. Lower panel, sideward view of the proposed oligomer; the membrane-binding sites are facing all in the same direction towards the putative membrane interface. Movies are available at http://www.endocytosis.org/EHDs/.

consistent with this interface being involved in oligomerization-dependent reorientation of residues responsible for ATP hydrolysis. Using the dimeric GBP1•GDP•AlF_3 G-domain structure[15] as a template and taking into account the observed 20-nm ring of the EHD2 coat (Fig. 1d), we predicted the arrangement of the EHD2 dimers within the oligomer, in which the nucleotides of two opposing EHD monomers are facing each other in a head-to-head fashion (see Methods and Fig. 4c). The predicted oligomer has a compact structure with a high degree of shape-complementarity between the oligomerizing dimers (Supplementary coordinates). Furthermore, the membrane-binding sites are pointing in the same direction towards the putative membrane interface (Fig. 4c, Supplementary Fig. 12a), and the thickness of the oligomeric ring agrees well with the thickness of the rings observed by electron microscopy (≈10 nm, Fig. 1d). In the predicted oligomer, the highly curved membrane interface of the EHD2 dimer is oriented perpendicular to the direction of the tubule curvature (Supplementary Fig. 12b, c). Furthermore, the disordered surface loop at the side of the G-domain containing the two conserved PF motifs (NPF and KPF) comes into the vicinity of the EH-domain linker (Fig. 4c and Supplementary Fig. 7b). Thus, we speculated that the EH-domain might switch position from the observed top site in the dimer in solution to the side-site position of the opposing dimer, during oligomerization (Fig. 4c). Interestingly, only a side-site NPF peptide, and not the top-site GPF peptide, bound with a measurable affinity (≈130 μM) to the isolated EH-domain of EHD2 (Supplementary Fig. 7c). Furthermore, in agreement with our prediction we observed that a deletion mutant of the EH-domain (ΔEH) or a double mutant of the two side-site xPF motifs (F122A/F128A) did not show any membrane-stimulated ATPase activity (Fig. 4b), and we did not find any regular oligomers on liposomes in electron microscopy studies for these mutants (see Supplementary Fig. 7 for further discussion).

On the basis of the conservation of structural and mechanistic elements we propose to expand the dynamin superfamily to include other multidomain large G-domain proteins such as EHDs and BDLPs. The structure outlined above, the architecture of the membrane interaction site and the proposed oligomer provide a framework to understand membrane remodelling for the EHD family, and perhaps for dynamin superfamily members. The EHD2 dimer interacts with membranes via ionic interactions mediated by a highly curved interface and we predict that this interaction results in the insertions of V321 and F322 into the hydrophobic phase of the lipid bilayer. Both the curved interface and the hydrophobic residue insertions (see synaptotagmin[21]) will result in the bending of the membrane towards the EHD2 dimer (buckling). The membrane curvature imposed by our proposed oligomer would be perpendicular to the curvature imposed by the concave membrane-binding face of the EHD2 dimer (Supplementary Fig. 12b). This would cause considerable curvature stress in the lipid bilayer and would thus facilitate the lipid rearrangement required for the formation of intermediate stages towards membrane fission/fusion[22]. Finally we observe that nucleotide hydrolysis is most probably leading to membrane scission *in vivo*, and thus we would argue that conformational changes induced by nucleotide hydrolysis are transmitted through helix 8 to the membrane-binding interface leading to further membrane destabilization.

METHODS SUMMARY

Isothermal titration calorimetry measurements were performed at 10 °C in 20 mM HEPES (pH 7.5), 300 mM NaCl, 2 mM $MgCl_2$. Liposome binding assays were performed using 0.33 mg ml^{-1} of 0.8-μm-filtered liposomes as described previously (www.endocytosis.org). For electron microscopic studies, 2.5 μM EHD2 in 20 mM HEPES (pH 7.5), 150 mM NaCl, 1 mM $MgCl_2$ was incubated for 15 min at 25 °C in the presence of 1 mM nucleotide and 0.05 mg ml^{-1} (final concentration) of the indicated 0.8-μm-filtered liposomes. Samples were spotted on carbon-coated copper grids (Canemco and Marivac) and negatively stained with 2% uranyl acetate.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 June; accepted 15 August 2007.
Published online 3 October 2007; corrected 18 October 2007.

1. Praefcke, G. J. & McMahon, H. T. The dynamin superfamily: universal membrane tubulation and fission molecules? *Nature Rev. Mol. Cell Biol.* **5,** 133–147 (2004).
2. Shao, Y. *et al.* Pincher, a pinocytic chaperone for nerve growth factor/TrkA signaling endosomes. *J. Cell Biol.* **157,** 679–691 (2002).
3. Grant, B. *et al.* Evidence that RME-1, a conserved *C. elegans* EH-domain protein, functions in endocytic recycling. *Nature Cell Biol.* **3,** 573–579 (2001).
4. Caplan, S. *et al.* A tubular EHD1-containing compartment involved in the recycling of major histocompatibility complex class I molecules to the plasma membrane. *EMBO J.* **21,** 2557–2567 (2002).
5. Blume, J. J., Halbach, A., Behrendt, D., Paulsson, M. & Plomann, M. EHD proteins are associated with tubular and vesicular compartments and interact with specific phospholipids. *Exp. Cell Res.* **313,** 219–231 (2007).
6. George, M. *et al.* Shared as well as distinct roles of EHD proteins revealed by biochemical and functional comparisons in mammalian cells and *C. elegans. BMC Cell Biol.* **8,** 3 (2007).
7. Jovic, M., Naslavsky, N., Rapaport, D., Horowitz, M. & Caplan, S. EHD1 regulates β1 integrin endosomal transport: effects on focal adhesions, cell spreading and migration. *J. Cell Sci.* **120,** 802–814 (2007).
8. Lin, S. X., Grant, B., Hirsh, D. & Maxfield, F. R. Rme-1 regulates the distribution and function of the endocytic recycling compartment in mammalian cells. *Nature Cell Biol.* **3,** 567–572 (2001).
9. Park, S. Y. *et al.* EHD2 interacts with the insulin-responsive glucose transporter (GLUT4) in rat adipocytes and may participate in insulin-induced GLUT4 recruitment. *Biochemistry* **43,** 7552–7562 (2004).
10. Rotem-Yehudar, R., Galperin, E. & Horowitz, M. Association of insulin-like growth factor 1 receptor with EHD1 and SNAP29. *J. Biol. Chem.* **276,** 33054–33060 (2001).
11. Valdez, G. *et al.* Pincher-mediated macroendocytosis underlies retrograde signaling by neurotrophin receptors. *J. Neurosci.* **25,** 5236–5247 (2005).
12. Braun, A. *et al.* EHD proteins associate with syndapin I and II and such interactions play a crucial role in endosomal recycling. *Mol. Biol. Cell* **16,** 3642–3658 (2005).
13. Lee, D. W. *et al.* ATP binding regulates oligomerization and endosome association of RME-1 family proteins. *J. Biol. Chem.* **280,** 17213–17220 (2005).
14. Guilherme, A. *et al.* EHD2 and the novel EH domain binding protein EHBP1 couple endocytosis to the actin cytoskeleton. *J. Biol. Chem.* **279,** 10593–10605 (2004).
15. Ghosh, A., Praefcke, G. J., Renault, L., Wittinghofer, A. & Herrmann, C. How guanylate-binding proteins achieve assembly-stimulated processive cleavage of GTP to GMP. *Nature* **440,** 101–104 (2006).
16. Reubold, T. F. *et al.* Crystal structure of the GTPase domain of rat dynamin 1. *Proc. Natl Acad. Sci. USA* **102,** 13093–13098 (2005).
17. Low, H. H. & Lowe, J. A bacterial dynamin-like protein. *Nature* **444,** 766–769 (2006).
18. de Beer, T., Carter, R. E., Lobel-Rice, K. E., Sorkin, A. & Overduin, M. Structure and Asn-Pro-Phe binding pocket of the Eps15 homology domain. *Science* **281,** 1357–1360 (1998).
19. de Beer, T. *et al.* Molecular mechanism of NPF recognition by EH domains. *Nature Struct. Biol.* **7,** 1018–1022 (2000).
20. Marks, B. *et al.* GTPase activity of dynamin and resulting conformation change are essential for endocytosis. *Nature* **410,** 231–235 (2001).
21. Martens, S., Kozlov, M. M. & McMahon, H. T. How synaptotagmin promotes membrane fusion. *Science* **316,** 1205–1208 (2007).
22. Zimmerberg, J. & Kozlov, M. M. How proteins produce cellular membrane curvature. *Nature Rev. Mol. Cell Biol.* **7,** 9–19 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Long-term fellowships supported O.D. (The International Human Frontier Science Program Organization), R.L. (Swedish Research Council) and S.M. (EMBO). We thank M. Plomann for providing the complementary DNAs for mammalian EHDs, and the ESRF beam staff in Grenoble for their support. The authors declare no competing financial interests.

Author Information The atomic coordinates of mouse EHD2 have been deposited in the Protein Data Bank (PDB) with the accession number 2QPT. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.T.McM. (hmm@mrc-lmb.cam.ac.uk) or O.D. (oliver.daumke@mdc-berlin.de).

METHODS

Protein expression and structure determination. Mouse EHD2 full-length protein, the ΔEH-domain construct (amino acids 1–404) and all mutants were expressed as N-terminal His-fusions followed by a PreScission cleavage site in *Escherichia coli* BL21 DE3 Rosetta (Novagen) from a modified pET28 vector. Bacterial cultures in TB medium were induced at a OD_{600} of 0.2 with 40 μM IPTG, and grown overnight at 18 °C. Bacteria were lysed in 50 mM HEPES (pH 7.5), 400 mM NaCl, 25 mM imidazole, 2.5 mM β-mercaptoethanol, 500 μM Pefablock SC (Boehringer Ingelheim) using an Emulsiflex homogenizer (Avestin). After centrifugation at 100,000g for 45 min at 4 °C, the soluble extract was applied to a NiNTA-column (Qiagen) equilibrated with lysis buffer. The column was extensively washed with 20 mM HEPES (pH 7.5), 700 mM NaCl, 30 mM imidazole, 2.5 mM β-mercaptoethanol, 1 mM ATP, 10 mM KCl, and shortly with 20 mM HEPES (pH 7.5), 300 mM NaCl, 25 mM imidazole, 2.5 mM β-mercaptoethanol. Bound protein was eluted with 20 mM HEPES (pH 7.5), 300 mM NaCl, 100 mM imidazole, 2.5 mM β-mercaptoethanol, and dialysed overnight at 4 °C against 20 mM HEPES (pH 7.5), 300 mM NaCl, 2.5 mM β-mercaptoethanol in the presence of 250 μg PreScission protease to cleave the His-tag. The protein was re-applied to a NiNTA column to which it bound under these buffer conditions in the absence of the His-tag. The column was extensively washed with 20 mM HEPES, 300 mM NaCl, 2.5 mM β-mercaptoethanol, and the protein finally eluted with 20 mM HEPES, 300 mM NaCl, 2.5 mM β-mercaptoethanol, 25 mM imidazole, concentrated and further purified using a Sephadex200 size-exclusion column (two consecutive runs for proteins used for the ATPase assays). Typical yields were 4 mg of purified EHD2/1 bacterial culture. At 300 mM NaCl, we could concentrate the protein to 40 mg ml^{-1}, but at lower salt concentration we observed some precipitation at this protein concentration. The protein was partially stabilized by 1 mM $MgCl_2$.

ATPase assays. Multiple turnover ATPase assays were performed in 20 mM HEPES (pH 7.5), 135 mM NaCl, 15 mM KCl, 1 mM $MgCl_2$ at 30 °C with 10 μM EHD2 (or mutants) as enzyme and 100 μM ATP as substrate, in the absence or presence of 1 mg ml^{-1} sonicated Folch (Sigma-Aldrich) liposomes, with an average diameter of 135 nm, as determined by dynamic light scattering. Reactions were started by the addition of the protein to the final reaction mix and nucleotide hydrolysis was followed using standard HPLC measurement[23]. Initial rates were determined by applying a linear fit to data points up to 40% nucleotide hydrolysis. For the dynamin reaction, 1 μM of protein with 1 mM of GTP as substrate has been used.

Crystallization and structure determination. For crystallization, a selenomethionine-substituted point mutant Q410A was prepared, as described[24]. This mutant showed identical biochemical properties as the wild-type protein but displayed less degradation in the linker region when incubated over longer periods of time. The protein was concentrated to 40 mg ml^{-1} and supplemented with 4 mM $MgCl_2$, 2 mM AMP-PNP (Sigma-Aldrich; both final concentrations). The hanging-drop vapour-diffusion method was used for crystallization. Protein solution (2 μl) was mixed with an equal volume of reservoir solution containing 3% PEG2000 MME, 50 mM MES (pH 6.4), 4 mM $MgCl_2$. Crystals appeared after one week at 4 °C and had dimensions of $\approx 0.2 \times 0.2 \times 0.05$ mm^3. For flash-freezing in liquid nitrogen, they were first transferred for 10 s into 50 mM MES (pH 6.4), 75 mM NaCl, 4 mM $MgCl_2$, 2 mM AMP-PNP, 14% MPD before incubation in the final cryo-solution containing 50 mM MES (pH 6.4), 75 mM NaCl, 4 mM $MgCl_2$, 2 mM AMP-PNP, 27% MPD. No crystals were obtained in the presence of ADP or in nucleotide-free conditions.

One data set at the selenium peak wavelength was collected from a single crystal at the ESRF beamline ID14-EH4 (see Supplementary Table 1) and processed and scaled using the Xds program suite[25]. Crystals belonged to the monoclinic crystal system and contained one molecule in the asymmetric unit. Thirteen out of sixteen selenium atoms were found from SHELXD[26] using the anomalous signal of the data set. Selenium sites were refined and initial phases were calculated using the program SHARP[27]. In the resulting electron density, the main chain was clearly traceable, and an initial model could be built using the XtalView package[28]. The model was refined using Refmac5 (ref. 29) with 3 TLS groups (Table 1). The asymmetric unit contains 477 amino acids, one AMP-PNP, one magnesium, one calcium and five water molecules and has an excellent geometry with all residues in the favoured and most favoured region of the Ramachandran plot as judged by the program Procheck[30]. Ribbon plots were prepared using the program Molscript[31] and rendered with Raster3D[32]. Surface conservation plots were prepared using the ConSurf server[33] and Ccp4 molecular graphics[34]. Electron potential maps were generated using Ccp4 molecular graphics. All other surface representations were prepared using Pymol[35]. To predict the arrangement of the EHD2 dimer in the oligomer, two EHD2 dimers were superimposed with one of the two monomers of the GBP1•GDP•AlF_3-dimer (PDB code 2B92) using Swisspdb viewer[36]. The EHD2 dimers were manually re-aligned to avoid amino acid clashes such that the two-fold axis between the oligomerizing EHD2 monomers was maintained. A high degree of shape complementarity between the EHD2 dimers in the resulting tetramer supported this approach (Supplementary Fig. 11). Furthermore, the lipid-binding sites of both EHD2 dimers are expected to contact the membrane, and this restraint is fulfilled in the tetramer. To obtain a 20 nm ring, an 18° tilt was introduced between the dimers. We refrained from energy-minimising of this structure because major conformational changes in the interface are expected to take place on oligomerization (ordering of switch I and switch II) and the resolution of the structure is not appropriate for an accurate prediction. The programs Superpose and Pdbset from Ccp4 (ref. 37) were used to generate the oligomer from the tetramer. PDB coordinates of the proposed oligomer are found in the Supplementary Materials. Additional movies of the EHD structure will be posted on http://www.endocytosis.org/EHDs/.

Ultracentrifugation. Sedimentation velocity experiments were performed in a Beckman Optima XLA ultracentrifuge, using an An-60Ti rotor. Centrifugation was at 50,000 revolutions min^{-1} and 5 °C at an EHD2 concentration of 15 μM, with scans as fast as possible (~1.5 min intervals). The data were analysed using DCDT+ v.2 (refs 38, 39), with the partial specific volume for the protein (from the amino acid composition) and solvent density and viscosity calculated using Sednterp[40]. Selected scans (at equal, ~15 min intervals), of $g(s_{20,w})$ (the amount of material sedimenting between $s_{20,w}$ and $s_{20,w}+\delta s$) and also the residuals for fitting the data with DCDT+, were plotted with the program Profit v.5.6.7 (Quantum soft).

Cell biology. Amino-terminal EGFP-tagged EHD2 and all mutants were overexpressed in HeLa cells from the pEGFP-C3 vector (Clontech). HeLa cells were grown in DMEM containing 10% fetal bovine serum and transfected using Genejuice (Novagen) for transient protein expression. Twenty four hours after transfection, cells were fixed for 20 min at 37 °C in 3.2% paraformaldehyde and mounted. All confocal images were taken sequentially using a BioRad Radiance system and LaserSharp software (Biorad). For real-time microscopy, transfected cells on glass-bottom Petri dishes (WillCo Wells BV) were washed with 25 mM HEPES (pH 7.5), 125 mM NaCl, 5 mM KCl, 10 mM D-glucose, 1 mM $MgCl_2$, 2 mM $CaCl_2$, and epifluorescence images were taken using an Olympus IX70 microscope (Southhall) and Argon laser (Melles Griot) with a Princeton instruments (Trenton)-cooled I-PentaMAX camera with MetaMorph software (Universal imaging).

23. Lenzen, C., Cool, R. H. & Wittinghofer, A. Analysis of intrinsic and CDC25-stimulated guanine nucleotide exchange of p21ras-nucleotide complexes by fluorescence measurements. *Methods Enzymol.* **255,** 95–109 (1995).
24. Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J. Mol. Biol.* **229,** 105–124 (1993).
25. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell vonstants. *J. of Appl. Crystallogr.* **26,** 795–800 (1993).
26. Sheldrick, G. M. & Schneider, T. R. SHELXL: High-resolution refinement. *Methods Enzymol.* **277,** 319–343 (1997).
27. de la Fortelle, E. & Bricogne, G. in *Methods in Enzymology* (eds Carter, C. W. Jr & Sweet, R. M.) 472–494 (1997).
28. McRee, D. E. XtalView/Xfit—A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125,** 156–65 (1999).
29. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240 (1997).
30. Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. Procheck—a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).
31. Kraulis, P. J. Molscript—a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24,** 946–950 (1991).
32. Merritt, E. A. & Murphy, M. E. Raster3D Version 2.0. A program for photorealistic molecular graphics. *Acta Crystallogr. D* **50,** 869–73 (1994).
33. Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33,** W299–W302 (2005).
34. Potterton, E., McNicholas, S., Krissinel, E., Cowtan, K. & Noble, M. The CCP4 molecular graphics project. *Acta Crystallogr. D* **58,** 1955–1957 (2002).
35. DeLano, W. L. *The PyMOL Molecular Graphics System* (DeLano Scientific, Palo Alto, California, USA, 2002).
36. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophor.* **18,** 2714–2723 (1997).
37. Collaborative Computational Project. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760 (1994).
38. Philo, J. S. A method for directly fitting the time derivative of sedimentation velocity data and an alternative algorithm for calculating sedimentation coefficient distribution functions. *Analyt. Biochem.* **279,** 151–163 (2000).

39. Philo, J. S. Improved methods for fitting sedimentation coefficient distributions derived by time-derivative techniques. *Analyt. Biochem.* **354,** 238–246 (2006).
40. Laue, T. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. in *Analytical Ultracentrifugation in Biochemistry and Polymer Science* (eds Harding, S. E., Rowe, A. J. & Horton, J. C.) 90–125 (Roy. Soc. of Chem., Cambridge, UK, 1992).

LETTERS

Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation

Antonis Kirmizis[1], Helena Santos-Rosa[1], Christopher J. Penkett[2], Michael A. Singer[3], Michiel Vermeulen[4], Matthias Mann[4], Jürg Bähler[2], Roland D. Green[3] & Tony Kouzarides[1]

Modifications on histones control important biological processes through their effects on chromatin structure[1–3]. Methylation at lysine 4 on histone H3 (H3K4) is found at the 5′ end of active genes and contributes to transcriptional activation by recruiting chromatin-remodelling enzymes[4,5]. An adjacent arginine residue (H3R2) is also known to be asymmetrically dimethylated (H3R2me2a) in mammalian cells[6], but its location within genes and its function in transcription are unknown. Here we show that H3R2 is also methylated in budding yeast (*Saccharomyces cerevisiae*), and by using an antibody specific for H3R2me2a in a chromatin immunoprecipitation-on-chip analysis we determine the distribution of this modification on the entire yeast genome. We find that H3R2me2a is enriched throughout all heterochromatic loci and inactive euchromatic genes and is present at the 3′ end of moderately transcribed genes. In all cases the pattern of H3R2 methylation is mutually exclusive with the trimethyl form of H3K4 (H3K4me3). We show that methylation at H3R2 abrogates the trimethylation of H3K4 by the Set1 methyltransferase. The specific effect on H3K4me3 results from the occlusion of Spp1, a Set1 methyltransferase subunit necessary for trimethylation. Thus, the inability of Spp1 to recognize H3 methylated at R2 prevents Set1 from trimethylating H3K4. These results provide the first mechanistic insight into the function of arginine methylation on chromatin.

Methylation at lysine and arginine residues within histones has been linked to gene expression[1–3]. Studies in mammalian cells have shown that arginine methylation of histones can influence both gene activation and repression. However, the precise mechanism employed by arginine methylation to exert its effects on the chromatin template is still unknown. In contrast, increasing evidence shows that lysine methylation modulates gene expression by recruiting downstream effector proteins. Recent findings showed that methylation at H3K4 (H3K4me) controls transcription activation by recruiting chromatin remodelling activities[4,5]. This recruitment can be specific for H3K4me3 (refs 7, 8), indicating that the three different methyl states of H3K4 (monomethylated, dimethylated and trimethylated) have distinct functions in gene expression. The Set1 complex is the enzyme that can mediate the methylation of H3K4, and various components of the complex regulate the establishment of the different methyl-H3K4 states[9–11].

To investigate the role of methylation at H3R2 in gene expression we raised an antibody against H3R2me2a. This modification is known to be catalysed by the mammalian CARM1/PRMT4 *in vitro*[12] and is affected by deletion of this methyltransferase in mouse embryonic fibroblasts[6]. Immunoblot analysis reveals that H3R2me2a is present *in vivo*, on mammalian and yeast histone H3 (Supplementary Fig. 1a, b). We confirmed the specificity of this antibody towards H3R2me2a by dot-blot analysis and peptide competition assays (Supplementary Fig. 1c, d). Most importantly, the antibody does not recognize histone H3 in yeast cells in which arginine 2 has been mutated to an alanine residue (H3R2A), a glutamine residue (H3R2Q) or a lysine residue (H3R2K) (Supplementary Figs 1e and 2).

To examine the specific function of H3R2 methylation, we used a high-resolution chromatin immunoprecipitation (ChIP)-on-chip analysis in yeast to determine its genome-wide distribution. We found that this mark is associated with both heterochromatin and euchromatin (Figs 1 and 2). Analysis of heterochromatin showed that H3R2me2a is present at all four heterochromatic regions in yeast: the two silent mating-type loci (*HMR* and *HML*), the ribosomal RNA-encoding DNA (*rDNA* repeat) and telomeres (Fig. 1a–d). In this analysis it became clear that all heterochromatic regions that were enriched in H3R2 methylation were also devoid of the active methyl mark H3K4me3 (Fig. 1a–d). Indeed, the decrease in H3R2me2a and increase in H3K4me3 levels can be used to define the boundaries of heterochromatic regions.

The presence of H3R2me2a at heterochromatic sites indicates that this methylation may be part of a signal to silence transcription. We therefore used yeast strains expressing the H3R2 mutants H3R2A and H3R2Q to test the role of this residue in heterochromatic silencing. We found that mutation of arginine 2 resulted in a severe loss of silencing in the *HMR*, telomere and *rDNA* loci, and to a moderate extent at the *HML* locus (Fig. 1e and Supplementary Fig. 3, left panels). These results indicate that arginine 2 on H3 is necessary for heterochromatic silencing, indicating a possible role for H3R2 methylation in this process.

We next considered the mechanism by which H3R2me2a may function to regulate heterochromatin. Comparison of the occupancy of key heterochromatic factors, such as Rap1p and Sir2p, with H3R2me2a enrichment indicates a coincidence at telomeric sites (compare Fig. 1d with Fig. 1f). However, when ChIP analysis of Rap1p and Sir2p was compared in wild-type (WT) and H3R2A strains the amount of binding of these two proteins at the heterochromatic sites was not changed (Fig. 1f). These results indicate that methylation at H3R2 functions at heterochromatin through a previously unobserved mechanism, which is independent of Rap1p and Sir2p recruitment. However, we cannot exclude the possibility that disruption of H3K4 methylation may contribute to the H3R2A phenotype (see below).

To determine the role of H3R2me2a within euchromatin, we divided 5,065 genes into five groups according to their transcriptional rate[13] (designated by shades of blue in Fig. 2a). We then examined the average enrichment of H3R2me2a for each gene group.

[1]Gurdon Institute and Department of Pathology, Tennis Court Road, Cambridge CB2 1QN, UK. [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK. [3]NimbleGen Systems, Inc., 1 Science Court, Madison, Wisconsin 53711, USA. [4]Max Planck Institute for Biochemistry, Department of Proteomics and Signal Transduction, D-82152 Martinsried, Germany.

Composite profiles indicate that H3R2me2a occurs near the middle of the coding region and peaks towards the 3′ end of genes (Fig. 2a). The H3R2me2a enrichment is inversely correlated with transcriptional activity because this mark is most abundant on the least active genes.

A role for H3R2 methylation in transcriptional repression was also highlighted by comparing its genomic profile with the pattern of H3K4me3, which is a signal for active transcription. Figure 2b shows that H3K4me3 is found at the 5′ end of genes, which is consistent with previous studies[14,15], and that it is most enriched at the most active genes (darker shades of blue). We noticed that H3K4me3 is enriched in the region of a gene at which H3R2me2a is missing, suggesting an antagonistic relationship between these two modifications (Fig. 2a, b).

To investigate the relationship of H3R2me2a with all H3K4 methyl states we divided all genes into three transcriptional categories (inactive, moderately transcribed and highly transcribed) and then compared the distribution of H3R2me2a with that of the three methyl-H3K4 marks (H3K4me1, H3K4me2 and H3K4me3) across individual genes. In all three transcriptional states the pattern of H3R2me2a was mutually exclusive with H3K4me3 specifically (Fig. 2c–e and Supplementary Fig. 4). This inverse enrichment between H3R2me2a and H3K4me3 is not seen with trimethylation of the other two known modified lysines in yeast, H3K36me3 and H3K79me3 (data not shown). The inverse profiles of H3R2me2a and H3K4me3 were not due to failure of the anti-H3R2me2a and anti-H3K4me3 antibodies to recognize their epitope when the adjacent residue is methylated (Supplementary Fig. 1c). These results indicate that H3R2me2a covers the promoter and coding region of silent genes but, as the transcription rate is increased, H3R2me2a recedes from the 5′ end and is replaced by H3K4 trimethylation.

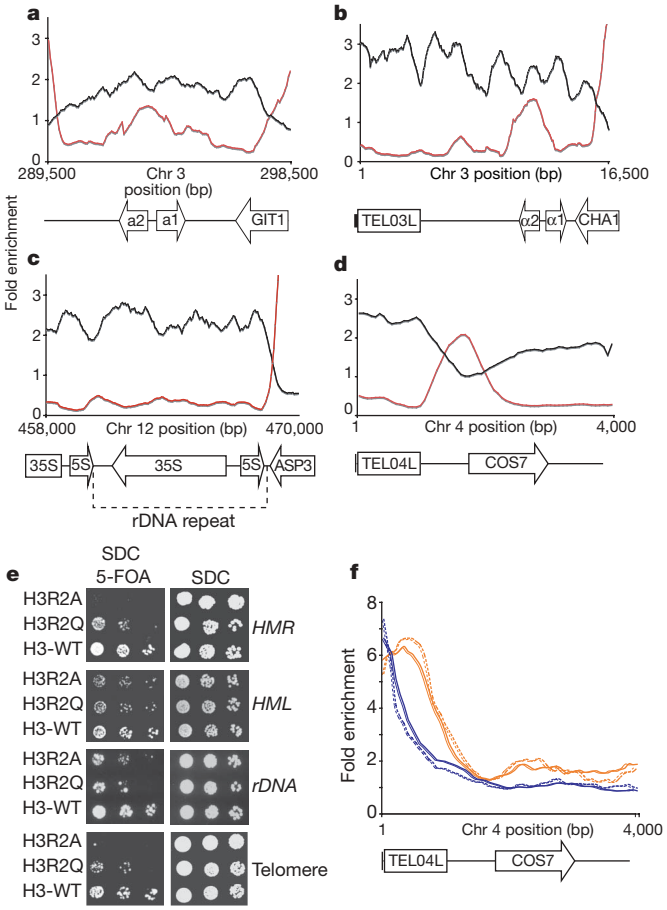


Figure 1 | H3R2me2a associates with heterochromatin. **a**–**d**, ChIP-on-chip analysis was performed in wild-type yeast cells (BY4741) grown to mid-exponential phase with the use of antibodies against H3R2me2a (black lines) and H3K4me3 (red lines). The graphs show a moving average (window = 15, step = 1) of the H3R2me2a and H3K4me3 enrichment normalized to histone H3 occupancy at the right (*HMR*, **a**) and left (*HML*, **b**) mating-type cassettes, at the silent ribosomal DNA repeat region (*rDNA*, **c**) and at the left telomere of chromosome 4 (*TEL04L*, **d**). Values less than 1 indicate regions that are not enriched. The arrows at the bottom of the graphs represent the locations of genes and the direction in which each is transcribed. The rectangles at the bottom of the graphs correspond to the telomeric sequences. bp, base pairs. **e**, Heterochromatin silencing assays were performed on cells from the H3R2A, H3R2Q and isogenic wild-type (H3-WT) yeast strains. Plates were photographed after incubation for 48 h at 30 °C. **f**, ChIP-on-chip analysis on telomeres with the use of antibodies against Sir2p (orange) and Rap1p (blue) in WT (solid lines) and H3R2A (dashed lines) strains. The graph shows a moving average (window = 15, step = 1) of the antibody enrichment over input at *TEL04L*.

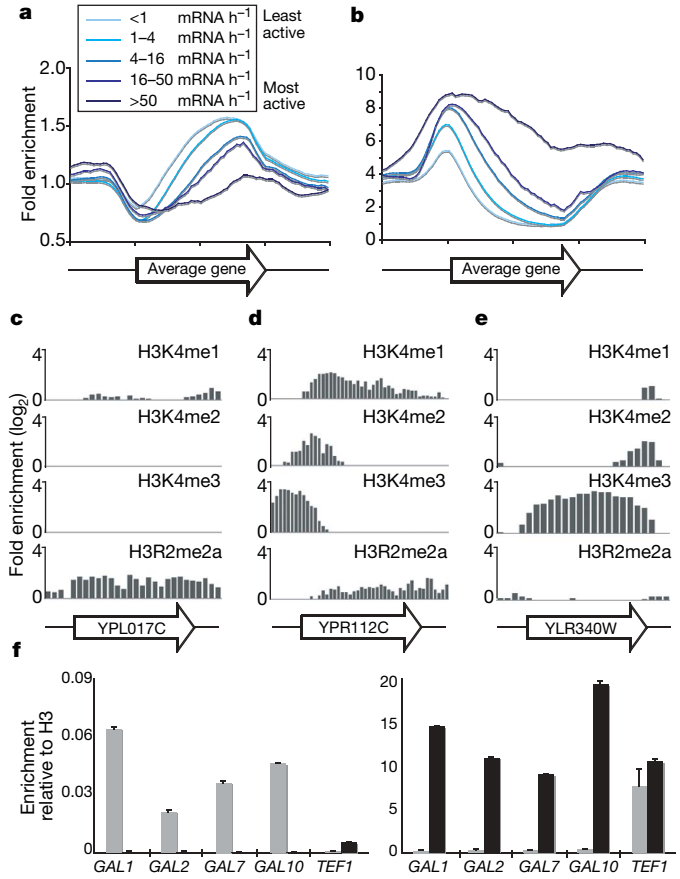


Figure 2 | H3R2me2a enrichment at euchromatic genes is mutually exclusive with H3K4me3. **a**, **b**, ChIP-on-chip analysis was performed in wild-type yeast cells (BY4741) grown to mid-exponential phase with the anti-H3R2me2a and anti-H3K4me3 antibodies. The graphs represent composite profiles of H3R2me2a (**a**) and H3K4me3 (**b**) at 5,065 genes, which were divided into five groups according to their transcriptional rate[12], shown as different shades of blue (defined in **a**). **c**–**e**, ChIP-on-chip analysis was performed as above, with antibodies against H3K4me1, H3K4me2, H3K4me3 and H3R2me2a. The distribution of these modifications is compared at three differentially expressed genes: an inactive gene (**c**), a moderately transcribed gene (**d**) and a highly transcribed gene (**e**). The name of each gene (arrow) is shown at the bottom of the graphs. **f**, ChIP experiments were performed in yeast cells grown in either glucose (repressed condition; grey bars) or galactose as carbon source (active condition; black bars) with antibodies against H3R2me2a (left) and H3K4me3 (right). The precipitated DNA was analysed by quantitative PCR (RT–PCR) with primers specific for the indicated genes. Error bars indicate s.e.m. for duplicate experiments. The induction of gene expression from glucose-containing to galactose-containing medium was monitored by reverse-transcriptase-mediated PCR analysis (see Supplementary Fig. 9).

We next examined whether H3R2me2a and H3K4me3 are dynamically exchanged on nucleosomes when gene expression is induced. ChIP analysis of cells grown in repressive conditions (glucose) showed high enrichment of H3R2me2a at *GAL* genes, whereas H3K4me3 was not detected at all on the same loci (Fig. 2f, grey bars). Shifting the cells to activating conditions (galactose) completely reversed the levels of the two modifications at those same locations (Fig. 2f, black bars). This analysis confirms the inverse correlation observed between H3R2me2a and H3K4me3, and shows that the two modifications are dynamically regulated together in the sense that when one mark is removed from nucleosomes the other is incorporated.

The dynamic exchange of these two modifications on nucleosomes suggested that the arginine residue at position 2 of histone H3 might have a direct function in regulating the methylation at the adjacent lysine 4. We examined this possibility by analysing the global methylation levels at H3K4 in yeast strains carrying mutations at H3R2 (H3R2A, H3R2Q and H3R2K). Figure 3a and Supplementary Fig. 2 show that in the H3R2A and H3R2Q strains the H3K4me3 signal is abolished, whereas in the H3R2K strain the H3K4me3 signal is greatly reduced. The H3K4me1 and H3K4me2 are, respectively, unaffected or very slightly reduced in these mutant strains (Fig. 3a, lane 3, and Supplementary Fig. 2, lanes 2–4). The disruption of H3K4me3 by mutating H3R2 is specific, because the H3R2A mutation does not affect the trimethyl state at the other methylated sites, namely H3K36 and H3K79 (Fig. 3a, lane 3). The loss of the H3K4me3 signal in the H3R2A mutant is not due to failure of the anti-H3K4me3 antibody to recognize its epitope when arginine 2 is mutated (Supplementary Fig. 5, lane 3, left and right panels).

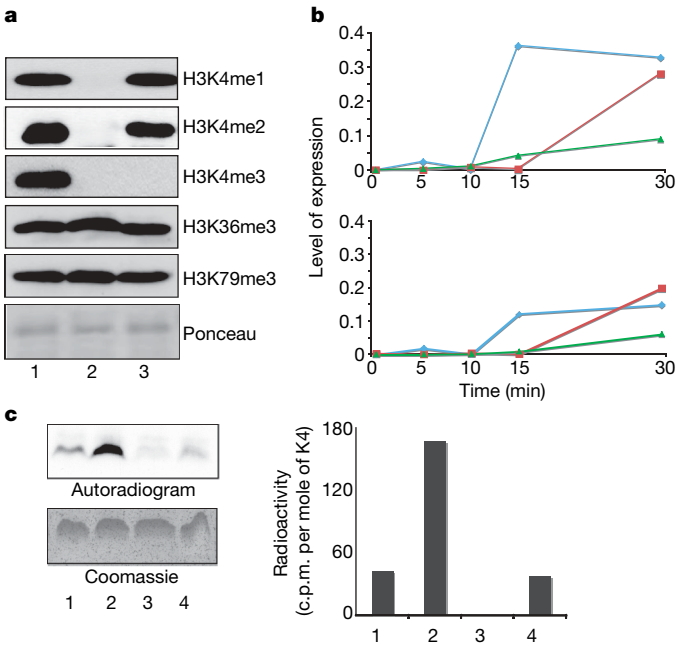


Figure 3 | H3R2me2a regulates the activity of the Set1 complex towards H3K4. a, Whole yeast extracts prepared from the H3R2A (lane 3), *set1C1068A* (lane 2) and isogenic WT (lane 1) strains were analysed by western blotting with antibodies against H3K4me1, H3K4me2, H3K4me3, H3K36me3 or H3K79me3. Uniform loading was monitored by Ponceau stain. **b**, Yeast cells expressing WT (blue), R2A mutant (red) or K4A mutant (green) histone H3 were grown to mid-exponential phase in medium containing raffinose (inactive condition) and then shifted to medium containing galactose (activating condition). RNA was prepared at the indicated times and analysed by quantitative RT–PCR with primers specific for *GAL7* (top) and *GAL1* (bottom). Values show the expression of *GAL7* or *GAL1* normalized to the RNA levels of *RTG2*, whose expression remains unchanged. The experiment was repeated with similar results. **c**, Purified Set1 complex from yeast was incubated with *S*-[^{3}H]adenosylmethionine in the presence of H3R2me2a (residues 1–17, lane 1), an unmodified peptide (residues 1–17, lane 2), H3K4me3 (residues 1–17, lane 3) or H3R2A mutant (residues 1–17, lane 4). Peptides were analysed for radioactive labelling with the use of autoradiography (activity on peptides; left) or sequential Edman degradation (activity on lysine 4; right). Equal loading of peptides was monitored by Coomassie stain (bottom left). The radioactivity released at lysine 4 of the H3K4me3 peptide (lane 3) was regarded as background signal and was subtracted from the lysine 4 signals of the other three peptides (right). The methylase assay was repeated three times with similar results.

The fact that trimethylation of H3K4 defines active transcription[16] prompted us to examine whether the H3R2A mutation affected gene expression. We tested the kinetics of *GAL1* and *GAL7* induction in the wild type and in H3R2A and H3K4A mutant strains. Figure 3b shows that mutation of H3R2 delays the activation of *GAL* genes. The effect of the H3R2A mutation is similar to, but less severe than, that of the H3K4A mutation, indicating that these two residues might be involved in a common regulatory mechanism (Fig. 3b).

We next sought to determine the mechanism responsible for the inverse distribution between H3R2me2 and H3K4me3. We first considered the possibility that H3R2 forms part of the recognition site for the Set1 complex, which methylates H3K4. Figure 3c shows that purified Set1 complex from yeast is able to methylate an unmodified H3 peptide but its activity is inhibited by mutation of arginine 2 to alanine (compare lane 2 with lane 4). Asymmetric dimethylation at H3R2 also inhibits the activity of the Set1 complex towards lysine 4 (Fig. 3c, compare lane 1 with lane 2). This Set1p activity is specific for lysine 4, because a peptide that is already trimethylated at H3K4 shows only background signal (Fig. 3c, lane 3). In addition, peptide sequencing reveals that the activity of the Set1 complex is occurring only at residue 4 (Supplementary Fig. 6). These results indicate that H3R2 is a recognition site for the Set1p methylase complex and also that methylation of H3R2 inhibits the Set1p enzyme from methylating H3K4.

The ability of Set1p to catalyse monomethylation, dimethylation or trimethylation at H3K4 is regulated by several components of the Set1 complex. Specifically, the Spp1 subunit is required for Set1-mediated trimethylation of H3K4 (ref. 10). It has been shown recently that Spp1 binds specifically to H3K4me2 and H3K4me3 through its PHD domain[17]. We therefore examined whether methylation of H3R2 is required for the binding of Spp1 to methylated lysine 4 *in vitro*. Figure 4a shows that dimethylation of H3R2 inhibits the interaction of the Spp1 PHD finger with dimethylated or trimethylated K4 (compare lanes 3 and 5 with lanes 4 and 6). Mutation of H3R2 also disrupts the binding of Spp1, which is consistent with the fact that this arginine residue is part of the recognition site of the Spp1 PHD finger[17] (Supplementary Fig. 7). To determine whether H3R2me2a also blocks Spp1 binding *in vivo* we performed ChIP assays in a yeast strain expressing Myc-tagged Spp1. Figure 4b and Supplementary Fig. 8 show that Spp1 is bound to regions of genes that were trimethylated at H3K4 and were devoid of H3R2me2a. However, Spp1 was absent from areas of genes at which H3R2 methylation was present, even though H3K4me1 or H3K4me2 was also abundant in these regions (Fig. 4b, middle panel, and Supplementary Fig. 8). These results confirm the biochemical analysis in Fig. 4a, which shows that H3R2 methylation can prevent Spp1 from binding H3K4me2. Moreover, the above observations that H3K4me3 is absent from regions in which H3R2me2a is enriched are consistent with the occlusion of Spp1.

Together these results identify the existence and indicate the function of H3R2 methylation in yeast. H3R2 methylation regulates the activity of the Set1 complex towards H3K4 by modulating the binding of its Spp1 component. The role of H3R2me2a in controlling H3K4me3 is also conserved in humans[18], indicating that this mechanism is likely to be generally applicable in all eukaryotes.

These findings place methylation at H3R2 and H3K4 in the same pathway and support a role of H3R2me2a as a negative regulator of H3K4 trimethylation. Figure 4c shows a model of how H3R2me2a may function during the transition from a repressed to a transcriptionally active state on a gene. Global analysis shows that when a gene

is inactive, H3R2me2a is present throughout the promoter and coding region (step 0). Methylation of H3R2 in yeast is likely to be catalysed by a previously unknown and as yet unidentified methyltransferase, because combinatorial deletion of the three known arginine methyltransferases (Rmt1, Rmt2 and Hsl7) does not affect the degree of this modification (data not shown). At this silent stage (step 0) very little, if any, methylation of H3K4 takes place. During activation, the presence of methylated H3R2 does not inhibit Set1p from monomethylating or dimethylating H3K4 (step 1). However, for trimethylation of H3K4 to take place, methylation at H3R2 has to be removed (step 2). The clearing of methylation at H3R2 must be mediated either by histone replacement or by the action of an as yet unidentified arginine demethylase. Once a region becomes devoid of H3R2 methylation, the Spp1 protein can recognize H3K4me2 by its PHD domain. This binding probably extends the time of interaction between the Set1 complex and its substrate, thus promoting the trimethylation of H3K4 by Set1p (step 3 (ref. 19)). Spp1 then associates with H3K4me3 (step 4), possibly to protect this methyl state from the action of the H3K4me3 demethylase Jhd2 (refs 20, 21). At the same time, Spp1 may protect H3R2 from methylation; structural studies[17,22,23] have shown that this arginine residue is absolutely required for the association of the Spp1 PHD finger with methylated H3K4. Together these data indicate that arginine methylation at H3R2 may influence transcription by regulating the H3K4 trimethylation capacity of the Set1 methyltransferase.

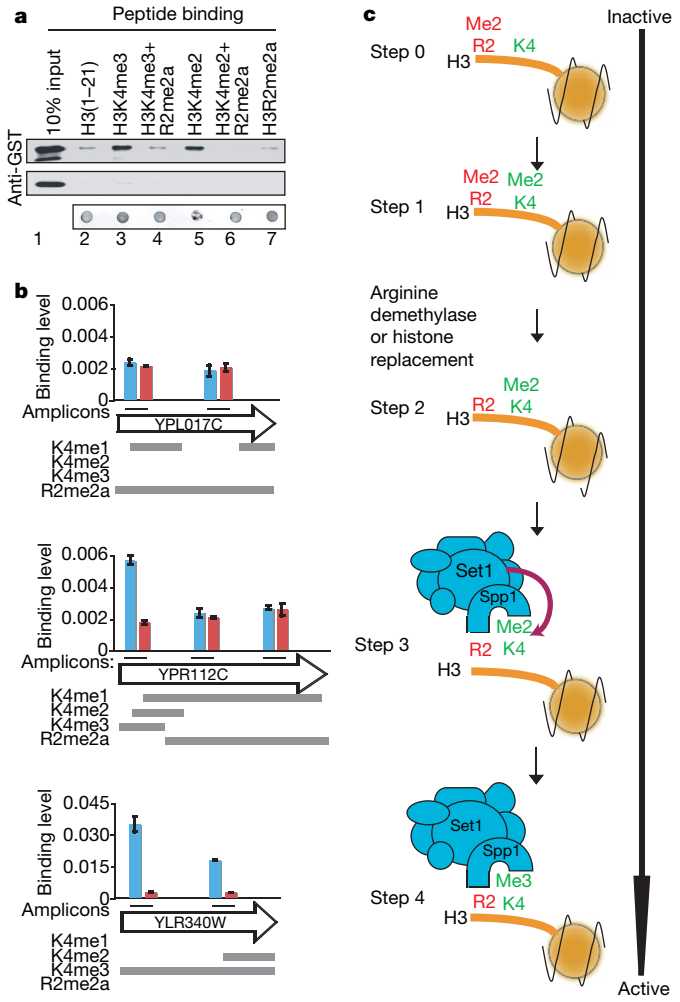


Figure 4 | H3R2me2a blocks the binding of Spp1 to methylated H3K4. **a**, *In vitro* binding assays were performed with synthetic N-terminal peptides (residues 1–21) of histone H3 and either recombinant glutathione *S*-transferase (GST)-conjugated $Spp1_{PHD}$ (top) or GST alone (middle). The bound proteins were monitored by western blot analysis with an anti-GST antibody. Peptide coupling was controlled by dot-blotting followed by Coomassie staining (bottom). **b**, *In vivo* binding analysis of Spp1 was performed with ChIP assays followed by quantitative PCR. Chromatin from yeast cells expressing a Myc-tagged (blue columns) or an untagged (red columns) form of Spp1 was immunoprecipitated with anti-Myc antibody. Three differentially expressed genes were examined: an inactive gene (top), a moderately transcribed gene (middle) and a highly transcribed gene (bottom). The analysed amplicons within each gene (named within the arrow) are indicated by black lines. Error bars indicate s.e.m. for duplicate experiments. The grey bars below each plot show the distribution of H3K4me1, H3K4me2, H3K4me3 and H3R2me2a within each gene. **c**, Model of how methylation at histone H3R2 controls trimethylation at H3K4.

METHODS SUMMARY

Formaldehyde crosslinking and ChIP were performed as described previously[24], with the following exceptions: the immunocomplexes were eluted from the Sepharose beads (17-5280-01; Amersham) using a total of 200 µl of elution buffer (100 mM sodium bicarbonate, 1% SDS), and treatment with RNase (11119915001; Roche) was performed during reversal of the crosslinks at 65 °C for 5 h. After reversal of the crosslinks, each individual ChIP sample was purified with the Qiaquick polymerase chain reaction (PCR) purification kit (Qiagen) and DNA was eluted from the columns with 50 µl of EB buffer (10 mM Tris-HCl pH 8.5). Amplicons were generated from individual ChIP samples by using a linker-mediated PCR. Sample labelling, hybridization and data extraction were performed by NimbleGen Systems Inc. as part of a ChIP Array Service. Downstream analysis of the ChIP-on-chip data was performed with the statistical package R (www.R-project.org) and associated array analysis modules in Bioconductor.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 March; accepted 9 August 2007.
Published online 26 September 2007.

1. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
2. Martin, C. & Zhang, Y. The diverse functions of histone lysine methylation. *Nature Rev. Mol. Cell Biol.* **6,** 838–849 (2005).
3. Wysocka, J., Allis, C. D. & Coonrod, S. Histone arginine methylation and its dynamic regulation. *Front. Biosci.* **11,** 344–355 (2006).
4. Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol. Cell* **25,** 15–30 (2007).
5. Sims, R. J. & Reinberg, D. Histone H3 Lys 4 methylation: caught in a bind? *Genes Dev.* **20,** 2779–2786 (2006).
6. Torres-Padilla, M. E., Parfitt, D. E., Kouzarides, T. & Zernicka-Goetz, M. Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* **445,** 214–218 (2007).
7. Shi, X. *et al.* ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* **442,** 96–99 (2006).
8. Wysocka, J. *et al.* A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442,** 86–90 (2006).
9. Dou, Y. *et al.* Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nature Struct. Mol. Biol.* **13,** 713–719 (2006).
10. Schneider, J. *et al.* Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression. *Mol. Cell* **19,** 849–856 (2005).
11. Steward, M. M. *et al.* Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nature Struct. Mol. Biol.* **13,** 852–854 (2006).
12. Schurter, B. T. *et al.* Methylation of histone H3 by coactivator-associated arginine methyltransferase 1. *Biochemistry* **40,** 5747–5756 (2001).
13. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728 (1998).
14. Liu, C. L. *et al.* Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol.* **3,** e328 (2005).
15. Pokholok, D. K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122,** 517–527 (2005).
16. Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419,** 407–411 (2002).
17. Shi, X. *et al.* Proteome-wide analysis in *Saccharomyces cerevisiae* identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36. *J. Biol. Chem.* **282,** 2450–2455 (2007).
18. Guccione, E. *et al.* Methylation of histone H3R2 by PRMT6 and H3K4 by an MLL complex are mutually exclusive. *Nature* doi:10.1038/nature06166 (this issue).
19. Wood, A. *et al.* Ctk complex-mediated regulation of histone methylation by COMPASS. *Mol. Cell. Biol.* **27,** 709–720 (2007).
20. Liang, G., Klose, R. J., Gardner, K. E. & Zhang, Y. Yeast Jhd2p is a histone H3 Lys4 trimethyl demethylase. *Nature Struct. Mol. Biol.* **14,** 243–245 (2007).

21. Seward, D. J. *et al.* Demethylation of trimethylated histone H3 Lys4 *in vivo* by JARID1 JmjC proteins. *Nature Struct. Mol. Biol.* **14,** 240–242 (2007).
22. Li, H. *et al.* Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442,** 91–95 (2006).
23. Pena, P. V. *et al.* Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* **442,** 100–103 (2006).
24. Morillon, A., O'Sullivan, J., Azad, A., Proudfoot, N. & Mellor, J. Regulation of elongating RNA polymerase II by forkhead transcription factors in yeast. *Science* **300,** 492–495 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Nelson, A. Bannister and M. Christophorou for critical reading of the manuscript; S. Marguerat for helpful discussions; L. Packman and M. Weldon for assistance with protein sequencing; N. Jiang and R. Selzer for assistance with microarray analyses; and M. Gilchrist for help with displaying genomic data. This work was supported by postdoctoral fellowship grants to A.K. from the European Molecular Biology Organization (EMBO) and Marie Curie. The T.K. laboratory is funded by grants from Cancer Research UK (CRUK) and the 6th Research Framework Programme of the European Union (Epitron and Heroic).

Author Information The microarray data sets are available from GEO (Gene Expression Omnibus) under accession number GSE8626, and from http://www.gurdon.cam.ac.uk/%7Ekouzarideslab/H3R2methylation.html. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to T.K. (t.kouzarides@gurdon.cam.ac.uk).

METHODS

Yeast strains and plasmids. The following *Saccharomyces cerevisiae* strains were used: wild-type (BY4741; Open Biosystems): *MAT***a**, *ura3-0*, *leu2Δ0*, *his3Δ1*, *met15Δ0*; JHY6: *MAT***a**, *ura3-52*, *lys2-801*, *ade2-101*, *trp*1−289, *his3Δ1*, *leu2*−3,112, Δ*hhf2-hht2*, Δ*hhf1-hht1*, pMS333[URA3-*HHT2-HHF2*]; UCC1369: *MAT***a**, *ade2::hisG*, *his3Δ200*, *leu2Δ0*, *lys2Δ0*, *met15Δ0*, *trp1Δ63*, *ura3Δ0*, *adh4::URA3-TEL-VIIL*, *ADE2-TEL-VR*, *hhf2-hht2::MET15*, *hhf1-hht1::LEU2*, pMP9; UCC1188: *MATα*, *leu2Δ1*, *lys2-801*, *trp1*, *ura3*, *RDN1::URA3*, *hhf2-hht2::HIS3*, *hhf1-hht1::LEU2*, pMP9; UCC7262: *MAT***a**, *ade2*, *his3*, *leu2*, *lys2*, *ura3*, *ADE2-TEL-VR*, *hhf2-hht2::MET15*, *hhf1-hht1::LEU2*, *hmra::URA3*, pMP9; UCC7266: *MAT***a**, *ade2*, *his3*, *leu2*, *lys2*, *ura3*, *ADE2-TEL-VR*, *hhf2-hht2::MET15*, *hhf1-hht1::LEU2*, *hml::URA3*, pMP9. C13 ABYS-86: *Matα ura3 leu2-3 his3-112 pra1-1 prb1-1 prc1-1 cps1-3*. Mutations of H3R2 and H3K4 were introduced into the *HHT2* gene in plasmid pMR206 (*TRP1-HHT2-HHF2*) using the QuickChange site-directed mutagenesis kit (Stratagene) and verified by sequencing. Mutant and wild-type control plasmids were introduced into strain JHY6 to make the strains JHY6-H3WT, JHY6-H3R2A, JHY6-H3R2Q and JHY6H3K4A. Derivatives of pMR206 were also used to replace plasmid pMP9 in strains UCC1369, UCC1188, UCC7262 and UCC7266 to make the following strains: UCC1369-H3WT, UCC1369-H3R2A, UCC1369-H3R2Q, UCC1188-H3WT, UCC1188-H3R2A, UCC1188-H3R2Q, UCC7262-H3WT, UCC7262-H3R2A, UCC7262-H3R2Q, UCC7266-H3WT, UCC7266-H3R2A and UCC7266-H3R2Q. Strain JHY6-H3WT was also used in PCR-mediated disruption to create strains JHY6-*set1C1068A* (*set1C1068A::HIS3*) and JHY6-*Sir2Δ* (*sir2::KAN*). Strain JHY6 was a gift from S. Berger, strain C13 ABYS-86 was a gift from D. H. Wolf, and strains UCC1369, UCC1188, UCC7262 and UCC7266 were gifts from D. Gottschling.

Western blotting. Yeast cells were grown to mid-exponential phase in rich medium in a 30 °C shaker. Total yeast extracts were prepared by first resuspending cell pellets in a tenfold volume of SDS loading buffer (3% 2-mercaptoethanol, 3% SDS, 0.1% bromophenol blue, 10% glycerol) and then the samples were alternately boiled and chilled three times to rupture cell membranes. Core calf thymus histones were obtained from Roche Applied Sciences. Acid extraction was used to prepare histones from H3K293 and C3H10T1/2 cells, which were cultured in DMEM medium supplemented with 10% FBS in a 5% CO_2 incubator at 37 °C. Western blotting was performed with standard procedures. The nitrocellulose membranes (Whatman) were blocked overnight in BSA buffer (Tris-buffered saline, 0.01% Tween 20, 5% BSA) at 4 °C. The antibodies used included anti-H3 (ab1791; Abcam), anti-H3R2me2a (ab8046; Abcam), anti-H3K4me1 (ab8895; Abcam), anti-H3K4me2 (ab7766; Abcam), anti-H3K4me3 (ab8580; Abcam), anti-H3K36me3 (ab9050; Abcam) and anti-H3K79me3 (ab2621; Abcam).

Antibody characterization. Dot-blotting was performed as described previously[25]. In brief, 2 μl of peptide dilutions were spotted on poly(vinylidene difluoride) (PVDF) membranes (RPN2020F; Amersham) and left to air-dry. The membranes were then blocked for 2 h with BSA buffer (Tris-buffered saline, 0.01% Tween 20 and 5% BSA) at room temperature (22–25 °C) before a standard immunoblot analysis was performed. For peptide competitions, 500 ng of isolated yeast histones, prepared as described below, were separated by 15% SDS–PAGE. A standard western blotting procedure was followed except that the antibodies were incubated with the peptides (0.5 μg ml^{-1}) for 30 min at room temperature before the membranes were probed. The antibodies used included anti-H3, anti-H3R2me2a and anti-H3K4me3.

The specificity of the anti-H3R2me2a antibody is further validated by the fact that its reactivity towards mammalian histones is sensitive to the deletion of CARM1/PRMT4 (ref. 6). This shows that the antibody is specific for methylated H3R2 because deletion of the enzyme (CARM1) that catalyses this modification *in vitro*[12] affects the antibody.

Yeast histone/chromatin preparation. Cells were grown to mid-exponential phase in rich medium (YPD), washed once in PBS and resuspended in Spheroplasting buffer (0.6 M sorbitol, 20 mM potassium phosphate pH 7.0, 5 mg of zymolase 20T per g of cells). After incubation for 30 min at room temperature with gentle shaking, cells were centrifuged at 3,500 r.p.m. for 10 min and washed once with 0.6 M sorbitol in 20 mM potassium phosphate pH 7.0). The cell pellet was then resuspended in 4 ml of Ficoll buffer (18% Ficoll, 20 mM PIPES pH 6.3, 0.5 mM $CaCl_2$, 1 mM dithiothreitol (DTT), 1 mM EDTA) per g of cells, disrupted by five or six strokes in a manual Dounce homogenizer and centrifuged at 6,000 r.p.m. for 10 min. The supernatant was discarded. Pelleted cells were resuspended in 10 ml of extraction buffer (10 mM HEPES pH 7.5, 1 mM EDTA, 0.5 M NaCl, 0.5% Nonidet P40), incubated on ice for 20 min and centrifuged at 14,000 r.p.m. for 10 min. The supernatant was discarded. The washes with extraction buffer were repeated until the pellet started to lose opacity. The pellet was then resuspended in 2 ml of Tris-HCl pH 8, 10% glycerol and dialysed overnight against the same buffer. Finally the sample was divided into aliquots and stored at −80 °C.

Chromatin immunoprecipitation. Wild-type (BY4741), JHY6-H3WT and JHY6-H3R2A strains were grown to mid-exponential phase in a 30 °C shaker. For the galactose induction experiments BY4741 cells were cultured overnight in complete medium (YPD) containing 2% glucose. The next morning the cells were split into two YPD samples, one containing 2% glucose and the other 2% galactose, and grown at 30 °C for a further 24 h. Cell pellets from both conditions collected at mid-exponential phase were used for ChIP experiments and RT–PCR (see below). The following antibodies were used for immunoprecipitation: 3 μl of anti-H3 per immunoprecipitation (IP), 3 μl/IP of anti-H3K4me3, 3 μl/IP of anti-H3K4me2, 3 μl per IP of anti-H3K4me1, 3 μl per IP of anti-H3R2me2a, 25 μl per IP of anti-Rap1 (sc-6663; Santa Cruz Biotechnology), 25 μl per IP of anti-Sir2 (sc-6667; Santa Cruz Biotechnology) and, as a negative control, 2 μl per IP of rabbit anti-mouse IgG (5180-2104; Biogenesis).

ChIP-on-chip data analysis. Primer positions were mapped with exonerate[26], and this information, along with an *S. cerevisiae* gff annotation file, was read into an R data structure with the use of scripts from the Bioconductor package tilingArray[27]. Raw NimbleGen output files were used for the analysis of the data. Two channels were used for each array; for each channel, the biweight mean for all data points was calculated and used to scale the data. The corrected data were then used to calculate a ratio of immunoprecipitated to control DNA for each spot. Finally, a ratio of these data to similarly corrected histone H3 data was used to normalize the data to nucleosome occupancy levels. Composite average profiles were created in a similar manner to that described previously[15]. The ends of open reading frames (ORFs) were defined at fixed points corresponding to the positions of translational start and stop sites. The length of the ORF was then subdivided into 40 regions of equal length, and the middle of each probe was assigned according to its nearest relative bin position. Probes 800 base pairs (bp) before start sites and 800 bp after stop sites were similarly assigned after subdivision into 20 bins for both regions.

Real-time PCR. Real-time PCR analysis was performed on an ABI PRISM 7000 sequence detection system with the use of SYBR Green (Applied Biosystems) as described previously[28]. In brief, standard curves for each primer set were calculated from amplification of wild-type genomic DNA diluted 1:10, 1:100. 1:1,000, 1:10,000 and 1:100,000. After each run, a dissociation curve was performed to ensure that no primer dimers contaminated the quantification and that the product had the expected melting temperature. Each PCR reaction was performed in duplicate and the analysis was repeated twice from independent ChIP experiments. A signal intensity value for each sample was calculated from the average of the two experiments. Relative fluorescent intensities for the ChIP experiments were calculated as follows: $[(\text{Ab signal}_X/\text{Ab signal}_Y) - (\text{IgG signal}_X/\text{IgG signal}_Y)]/[(\text{H3 signal}_X/\text{H3 signal}_Y) - (\text{IgG signal}_X/\text{IgG signal}_Y)]$, where Ab is the antibody of interest, IgG is the negative control antibody, H3 is the histone H3 antibody, X is the locus of interest and Y is the intergenic region on chromosome V that was used as an internal background control. The sequences of the primers used for PCR analysis are shown in Supplementary Table 1.

Quantitative RT–PCR. Total yeast RNA was prepared from 3×10^7 cells of each indicated growth condition using the RNeasy Mini kit (Qiagen) in accordance with the manufacturer's protocol. To ensure complete removal of contaminating DNA from the RNA preparations, the Turbo DNA-free kit (Ambion) was used. First-strand complementary DNA synthesis was achieved with SuperScript II reverse transcriptase (catalogue no. 18080; Invitrogen) with a primer cocktail containing 50 μM oligo(dT) (Ambion) and 50 ng of random hexamers (Invitrogen), as described in the manufacturer's instructions. The cDNA samples were then used as templates for real-time PCR (see above).

Array design. The *S. cerevisiae* genome tiling array contained a total of 379,521 50-mer oligonucleotides, positioned every 64 bp throughout the yeast genome representing both DNA strands. The design included random GC probes as controls.

Peptide synthesis. Peptides corresponding to amino-acid residues 1–17 of histone H3 were synthesized in house using the Fmoc strategy on a solid-phase peptide synthesizer (Intavis). Peptides were synthesized on amide resin as 20-mers containing the first 17 amino acids of the amino-terminal tail of histone H3 followed by a double glycine spacer and a biotinylated lysine residue on the carboxy terminus. Unmodified amino acids as well as methylated arginine and lysine derivatives used for the synthesis were purchased from Novagen or Bachem. After synthesis, peptides were cleaved from the resin with trifluoroacetic acid, precipitated with ether, and dried in air. The identity and quality of the peptides were checked by mass spectrometry. Histone H3 N-terminal peptides corresponding to residues 1–8 and 1–21 were obtained from Almac Sciences. Peptides were synthesized with at least 90% purity and were resuspended in 10 mM HEPES pH 7.5 containing 0.005% Igepal (Sigma).

***In vitro* methyltransferase and peptide pull-down assays.** Purification of the Set1 complex and the methyltransferase assay were performed as described[16]. Protein-A–Set1 was expressed and the Set1 complex was purified from the C13 ABYS-86 strain. About 1 μg of Set1 complex was incubated in 50 mM Tris-HCl pH 8.5, 20 mM KCl, 10 mM $MgCl_2$, 10 mM 2-mercaptoethanol, 0.05 mM DTT, 250 mM sucrose, 0.2% dodecyl-β-D-maltoside buffer, with 2 μl of *S*-adenosyl-L-[Me-^{3}H]methionine (TRK865; Amersham Pharmacia) and about 2 μg of peptides corresponding to residues 1–17 of histone H3 as substrate, for 1 h at 30 °C. The reactions were resolved in 17% Tricine gels, transferred by Western blotting to PVDF membrane and exposed to KodaK Biomax MS autoradiogram films for 12 h. Peptides in the reaction were also revealed by Coomassie staining of the PVDF membrane, cut off and subjected to Edman degradation at the PNAC facility, University of Cambridge (www.bioc.cam.ac.uk/pnac/proteinsequencing.html) followed by quantification in a scintillation counter (LS6500; Beckman Coulter). The molar concentration of each peptide was determined from the amount of the first alanine residue present in each peptide. The radioactive counts released from each peptide were then normalized to the molar concentration of each peptide.

The Spp1 PHD domain, covering amino acids 20–76, was amplified from genomic yeast DNA and cloned into vector pGEX-5X-1. Binding assays with 3 μg of GST–$Spp1_{PHD}$ and 20 μg of N-terminal H3 peptides were performed as described previously[29], with minor modifications. The binding was performed overnight at 4 °C in PDB-150 buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 10 μM $ZnCl_2$, 5 mM EDTA, 0.5% Igepal, containing protease inhibitor cocktail (Roche Applied Sciences)), followed by washing in PDB-150.

25. Perez-Burgos, L. *et al.* Generation and characterization of methyl-lysine histone antibodies. *Methods Enzymol.* **376,** 234–254 (2004).
26. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
27. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103,** 5320–5325 (2006).
28. Santos-Rosa, H. *et al.* Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. *Mol. Cell* **12,** 1325–1332 (2003).
29. Nelson, C. J., Santos-Rosa, H. & Kouzarides, T. Proline isomerization of histone H3 regulates lysine methylation and gene expression. *Cell* **126,** 905–916 (2006).

LETTERS

Methylation of histone H3R2 by PRMT6 and H3K4 by an MLL complex are mutually exclusive

Ernesto Guccione[1], Christian Bassi[1], Fabio Casadio[1], Francesca Martinato[1], Matteo Cesaroni[1], Henning Schuchlautz[2], Bernhard Lüscher[2] & Bruno Amati[1]

Eukaryotic genomes are organized into active (euchromatic) and inactive (heterochromatic) chromatin domains. Post-translational modifications of histones (or 'marks') are key in defining these functional states, particularly in promoter regions[1,2]. Mutual regulatory interactions between these marks—and the enzymes that catalyse them—contribute to the shaping of this epigenetic landscape, in a manner that remains to be fully elucidated[1,2]. We previously observed that asymmetric di-methylation of histone H3 arginine 2 (H3R2me2a) counter-correlates with di- and tri- methylation of H3 lysine 4 (H3K4me2, H3K4me3) on human promoters[3]. Here we show that the arginine methyltransferase PRMT6 catalyses H3R2 di-methylation *in vitro* and controls global levels of H3R2me2a *in vivo*. H3R2 methylation by PRMT6 was prevented by the presence of H3K4me3 on the H3 tail. Conversely, the H3R2me2a mark prevented methylation of H3K4 as well as binding to the H3 tail by an ASH2/WDR5/MLL-family methyltransferase complex[4–7]. Chromatin immunoprecipitation showed that H3R2me2a was distributed within the body and at the 3′ end of human genes, regardless of their transcriptional state, whereas it was selectively and locally depleted from active promoters, coincident with the presence of H3K4me3. Hence, the mutual antagonism between H3R2 and H3K4 methylation, together with the association of MLL-family complexes with the basal transcription machinery[8], may contribute to the localized patterns of H3K4 tri-methylation characteristic of transcriptionally poised or active promoters in mammalian genomes[1–3,9,10].

The amino-terminal tails of nucleosomal histones are rich in lysine and arginine, the nitrogen groups of which can be post-translationally modified by one or more methyl groups. Lysines may be mono (abbreviated me1), di (me2), or tri-methylated (me3), as well as acetylated or ubiquitinated[1], whereas arginines can be mono- or di-methylated, the latter either in a symmetric (me2s) or asymmetric form (me2a)[11,12]. Methylation has two main consequences: first, it creates possible binding sites for proteins with Chromo/Tudor-family domains[13] or PHD fingers[14]; second, it disrupts potential hydrogen bonding. These covalent histone modifications (or 'marks') are not randomly arranged on chromatin; characteristic combinations of marks are found at specific locations of the genome and are associated with different functional states. For example, H3K9me3 and H3K27me3 are repressive marks, whereas H3K4me2 and H3K4me3 are characteristic of active promoter regions, where they co-exist with additional methylation and acetylation marks[1,2]. We previously profiled 35 histone marks on the 5′ region of 151 human genes by quantitative chromatin immunoprecipitation (qChIP) and identified several new positive and negative associations[3]: among these, we noted a strong counter-correlation between H3K4me3 and H3R2me2a, which we characterize here in further detail.

We used qChIP to measure the levels of H3K4me3 and H3R2me2a on a collection of 110 human promoters in P493 cells[3] and of 75 promoters in HL-60 cells (Supplementary Fig. 1a, b). Most promoters enriched for H3K4me3 showed only background H3R2me2a levels, and vice-versa. This counter-correlation was not due to epitope masking, because each antibody recognized its cognate mark on an H3 tail peptide independently of the presence of the other mark (Supplementary Fig. 2). H3R2me2a also counter-correlated with messenger RNA levels, whereas H3K4me3 showed a strong positive correlation[3] (Supplementary Fig. 1c–f).

H3K4me3 is localized to short stretches that overlap with—or are proximal to—promoter regions[1–3,9,10]. To address the distribution of H3R2me2a, we designed PCR amplicons localized at the transcription start site (TSS), within the body and at the 3′ end of 44 human genes (representative maps shown in Supplementary Fig. 3). To visualize the data, the genes were ordered according to their mRNA expression levels in P493 cells (Fig. 1a) and aligned with the qChIP values (Fig. 1b–f). As expected, H3K4me3 was enriched in the promoter of expressed genes and absent from downstream sequences (Fig. 1b). H3R2me2a was highest on inactive promoters[3], but was also enriched inside and at the 3′ end of all the genes analysed, regardless of expression levels (Fig. 1c). This differed from the distribution of H3K27me3, which was present primarily at the 5′ end of repressed genes (Fig. 1d), or of H3K36me3, which was enriched inside active genes (Fig. 1e)[1,2]. Finer mapping at four transcribed loci with multiple PCR amplicons covering ~8 kb (−4 to +4 kb from the TSS) confirmed that H3R2me2a was depleted from the 5′ region, coincident with the presence of H3K4me3 (Supplementary Fig. 4). In summary, H3K4me3 and H3R2me2a are enriched on active and inactive promoters, respectively, whereas within genes H3K4me3 levels drop and H3R2me2a is found independently of transcriptional activity.

To characterize the relationship between H3K4me3 and H3R2me2a in further detail, we sought to identify the enzyme catalysing H3R2me2a. Arginines are methylated by protein arginine methyltransferases (PRMTs), with type I and II PRMTs catalysing asymmetric and symmetric di-methylation, respectively[11,12]. The human genome encodes 9 characterized PRMTs, 6 of which (PRMT1/8, 2, 3, 4 and 6) are type I enzymes. Substrate specificities are incompletely understood; PRMTs methylate many cellular proteins and not all may target histones. PRMT1 and PRMT4 (also called CARM1) introduce the H4R3me2a and H3R17/26me2a marks, respectively, which are both linked to transcriptional activation[11,12,15–20]. Albeit with lower efficiency, PRMT4 also methylates H3R2 *in vitro*[20]. qChIP analysis of H3R26me2a showed a distribution distinct from H3R2me2a, in particular at promoters, where H3R26me2a correlates with H3K4me3 and gene activity (Fig. 1f)[3]. Hence, it seemed unlikely that the same enzyme catalyses asymmetric di-methylation of R2 and R26.

[1]Department of Experimental Oncology, European Institute of Oncology (IEO), IFOM-IEO Campus, Milan 20139, Italy. [2]Institute of Biochemistry, Division of Biochemistry and Molecular Biology, RWTH Aachen University Hospital, 52074 Aachen, Germany.

We used small-interfering RNA (siRNA) oligonucleotides to downregulate the human *PRMT1–6* mRNAs in HeLa cells. The *PRMT6*-directed siRNA selectively caused reduction of bulk levels of H3R2me2a and, as expected, of the PRMT6 protein (Fig. 2a). Conversely, H3R2me2a levels in 293T cells were enhanced following overexpression of PRMT6, but not of a catalytically inactive mutant (PRMT6dn) or of PRMT4 (Fig. 2b). When immunoprecipitated from transfected cells and incubated with an H3 tail peptide, PRMT6 catalysed asymmetric di-methylation of H3R2, whereas PRMT6dn or PRMT4 failed to do so (Fig. 2c and data not shown). The presence of the H3K4me3 mark on the peptide prevented methylation of H3R2. Thus, PRMT6 is an H3R2 methyltransferase, whose ability to catalyse the H3R2me2a mark is precluded by prior deposition of H3K4me3.

The above data and the fact that recombinant PRMT4 preferentially targets H3R17/26 over H3R2 *in vitro*[20] prompt reconsideration of the relative roles of PRMT4 and PRMT6 in H3R2 methylation *in vivo*. Although PRMT4 did not seem to influence bulk H3R2me2a levels in HeLa cells (Fig. 2a, b), levels of H3R2me2a dropped in *PRMT4*-null mouse embryo fibroblasts[21]. Thus, PRMT4 and PRMT6 may target H3R2 in different cells, perhaps dependent on different expression levels or association with co-factors. Alternatively, deletion of PRMT4 may affect PRMT6 expression or activity, thereby indirectly altering H3R2 methylation. Further investigation is required to resolve this question.

Methyltransferases of the SET1/MLL family catalyse methylation of H3K4 and occur in evolutionarily conserved multi-subunit complexes, which include the proteins WDR5, ASH2 (also known as ASH2L) and menin (also known as MEN1)[4–7]. A complex purified through a TAP-tagged ASH2 protein methylated an H3 tail peptide, but not a peptide bearing the H3R2me2a mark (Fig. 3a). A similar result was obtained by immunoprecipitation of a Flag-tagged ASH2 protein (data not shown). Thus, H3R2me2a excludes H3K4 methylation by MLL complexes. Previous studies provide a molecular basis

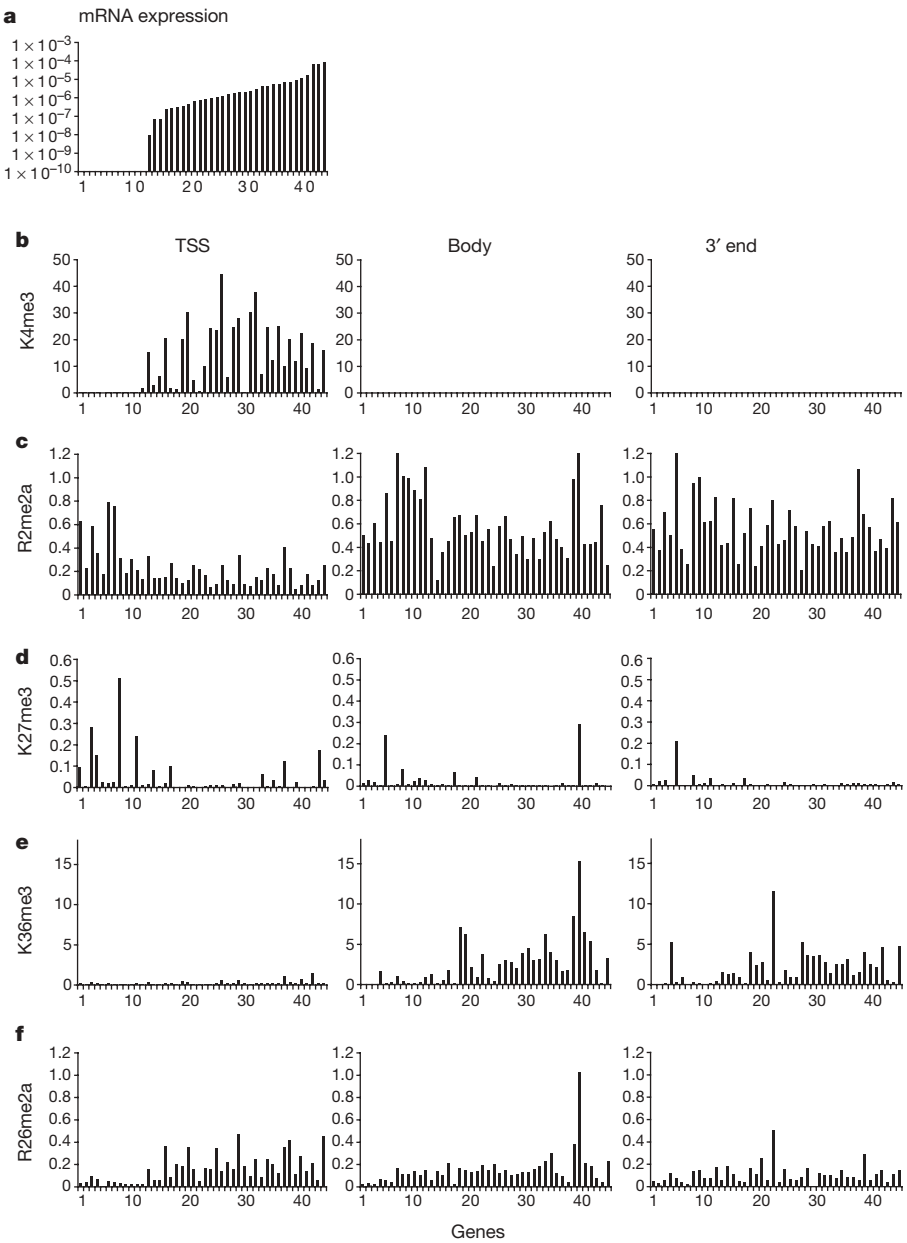


Figure 1 | H3R2me2a is present on inactive promoters, and within genes independently from the expression status. **a**, Genes were ordered on the basis of mRNA expression levels in P493 cells. qChIP values are shown in the same order for: **b**, H3K4me3; **c**, H3R2me2a; **d**, H3K27me3; **e**, H3K36me3; **f**, H3R26me2a. These marks were measured at the transcription start site (TSS), in the body or at the 3′ end of each gene, as indicated. Representative amplicon positions are shown in Supplementary Fig. 3. qChIP values are given as DNA recovery (%input) normalized for the %input of total H3.

for this finding: WDR5 recognizes the H3 tail independently of H3K4 methylation, mediates presentation of H3K4 to MLL, and is required for its methylation in cells[14,22–27]. WDR5 establishes molecular contacts with H3R2 (refs 23, 24), methylation of which interferes with WDR5 binding[23]. We confirmed this interaction in pull-down assays with H3 tail peptides bearing different modifications. WDR5 was recovered from U937 nuclear lysates with the unmodified H3 tail; the presence of H3K4me3 did not impinge on this interaction, but H3R2me2a reduced recovery to background levels (Fig. 3b). The same was observed with *in vitro*-translated WDR5 (Fig. 3c). As specificity controls, cellular HP1α was pulled down with an H3K9me3 peptide (data not shown), and the chromodomains of HP1β and CHD1 interacted with H3K9me3 and H3K4me3, respectively (Supplementary Fig. 5). Besides WDR5, H3R2me2a is likely to exclude binding of other H3 ligands, including PHD finger proteins that recognize K4-methylated forms of the H3 tail and form contacts with the unmodified R2 residue[14].

Given that H3R2me2a prevents interaction of WDR5 with the H3 tail, one would expect SET1/MLL complexes to be excluded from chromatin domains bearing this mark. qChIP analysis of WDR5 and ASH2 in HL60 cells showed their co-occurrence on a subset of promoters (Fig. 4a), as well as their correlation with H3K4me3 (data not shown) and counter-correlation with H3R2me2a (Fig. 4b,c). In P493 cells, WDR5 and menin also correlated with H3K4me3 (Fig. 4d)[3] and counter-correlated with H3R2me2a (Fig. 4e,f).

We have demonstrated a reciprocal interference between methylation of H3R2 and H3K4. The H3R2me2a mark prevents recognition of the H3 tail by WDR5, recruitment of MLL methyltransferase complexes and H3K4 methylation. Reciprocally, H3K4me3 prevents H3R2 methylation by PRMT6. The structural basis for H3 recognition by PRMT6 remains to be addressed: we detected no stable interaction between PRMT6 and H3 tail peptides, indicating that additional nucleosome domains or proteins might be involved. The antagonism between H3R2 and H3K4 methylation is likely to directly contribute to their mutually exclusive patterns in chromatin. H3K4me3 is selectively associated with transcriptionally poised or active promoters, most probably owing to the interaction of SET1/MLL complexes with RNA polymerase II[8]. H3R2me2a is present on inactive promoters, as well as in the body of all genes examined, regardless of expression levels: although the mechanistic basis for this deposition pattern remains to be unravelled, the selective depletion of H3R2me2a from active promoters is likely to be imposed by the presence of H3K4me3. On the other hand, H3K4me3 generally does not spread into the body of active genes in mammalian genomes[3,9,10]—we suggest that H3K4me3 spreading is restricted in part by the presence of H3R2me2a within genes. It is worth noting here that H3K4me2 frequently spreads further along into the body of transcribed genes[3,9]; because ASH2-associated MLL activity is most critical for H3K4 tri- as opposed to di-methylation (H.S. and B.L., manuscript in preparation)[7], it will be relevant to address whether other H3K4-specific methyltransferases[1] are insensitive to the presence of H3R2me2a. In spite of the antagonism between H3R2me2a

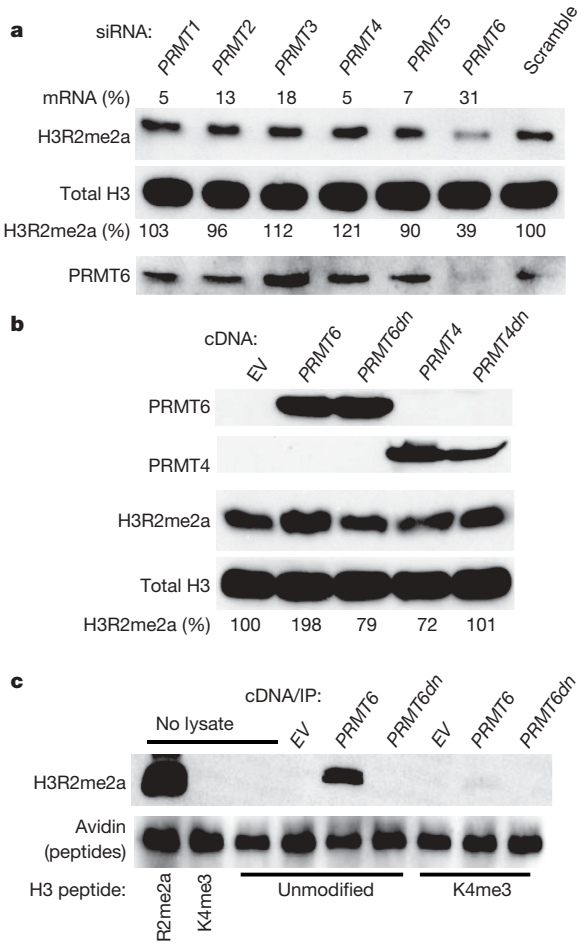


Figure 2 | H3R2 methylation by PRMT6 is prevented by H3K4me3. **a**, **b**, HeLa cells were transfected with siRNA oligonucleotides directed against the indicated PRMTs (**a**) or cDNA expression vectors encoding wild-type or mutant (*dn*) forms of PRMT6 or PRMT4 (bearing Myc- and HA-epitope tags, respectively) (**b**). Immunoblotting was used to analyse bulk levels of H3R2me2, total histone H3, endogenous PRMT6 (**a**) or epitope-tagged PRMT4/6 (**b**), as indicated on the left. mRNA (%): the residual level of the corresponding *PRMT* mRNA relative to that of the same mRNA in the non-specific siRNA sample (Scramble). EV, empty vector. H3R2me2a (%): H3R2me2a/total H3 ratio, determined by densitometric scanning and normalized to that in the control sample (Scramble in **a**, EV in **b**). **c**, Anti Myc-tag immunoprecipitates from transfected cells (as indicated above) were incubated with H3 tail peptides (indicated below).

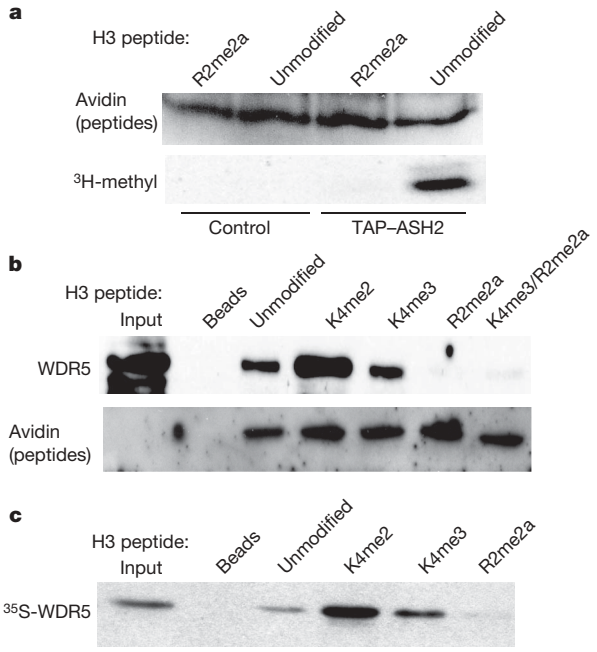


Figure 3 | H3R2me2a excludes binding of WDR5 to the H3 tail. **a**, Methylation of an H3 tail peptide by a purified TAP–ASH2 complex is sensitive to pre-methylation of H3R2. Biotinylated H3 tail peptides (indicated above) were incubated with the TAP–ASH2 complex or buffer only (control) in the presence of radioactive ^{3}H-Adomet and run on SDS–PAGE. Incorporation of the radioactive methyl group (^{3}H-methyl) was monitored by autoradiography. Avidin was used to visualize total peptide levels in each sample. **b**, The H3 tail peptides were incubated with U937 nuclear extracts, pulled down on avidin-conjugated beads, followed by SDS–PAGE and immunoblotting with anti-WDR5. **c**, Same as **b** with *in vitro*-translated, ^{35}S-labelled WDR5, visualized by autoradiography.

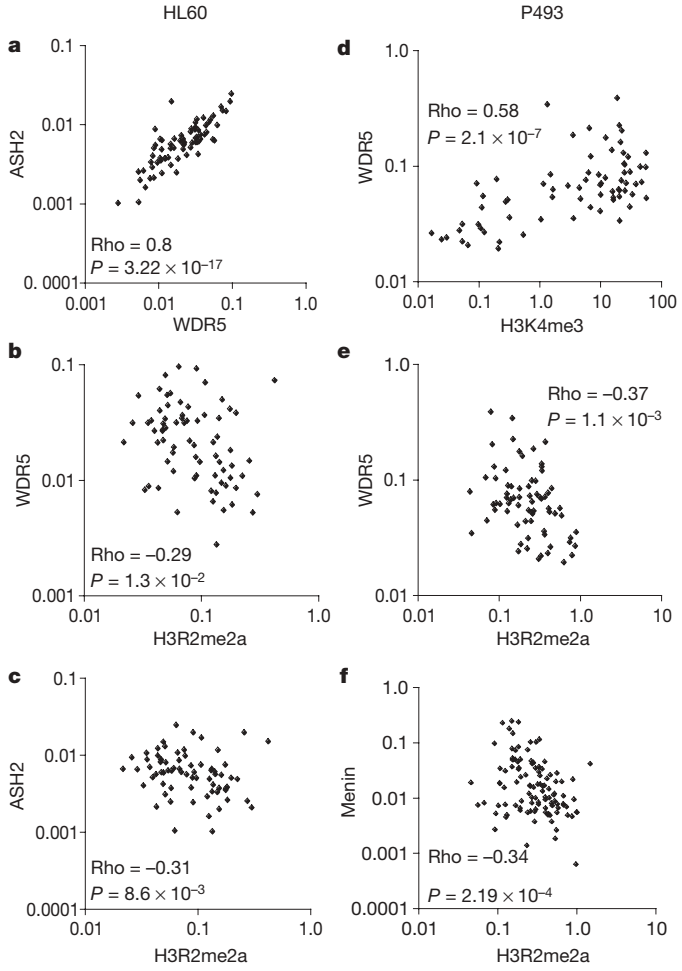


Figure 4 | H3R2me2a and MLL complex subunits are counter-correlated on chromatin. Two-by-two comparisons of qChIP values for the indicated proteins (ASH2, WDR5 and menin, expressed as percentage of input) and histone marks (H3R2me2a and H3K4me3, as percentage of input and normalized to total H3) in HL60 (**a**–**c**) and P493 cells (**d**–**f**). Each data point represents recovery of a given DNA site by qChIP. The data regard only promoter regions.

and H3K4me3 deposition, PRMT6 and ASH2/MLL activities seem to be interdependent at a general regulatory level, because siRNAs against either PRMT6 or ASH2 (which is critical for MLL activity in cells)[7] affected both H3R2me2a and H3K4me3 (Supplementary Fig. 6). This finding further emphasizes that the negative crosstalk between H3R2me2a and H3K4me3 occurs locally on chromatin. The exclusion of H3K4me3 deposition by H3R2me2a and the counter-correlation between those marks are evolutionarily conserved features observed also in the *Saccharomyces cerevisiae* genome[28].

In four-cell stage mouse embryos, pluripotent blastomeres that will give rise to the inner cell mass are marked by high levels of H3R2me2a, H3R17me2a and H3R26me2a. Most remarkably, PRMT4/CARM1 overexpression in blastomeres promoted inner cell mass formation[21]. These observations warrant elucidation of PRMT6's function and of its possible interplay with PRMT4 in H3R2 methylation and cell fate determination during early embryonic development.

METHODS SUMMARY

Growing P493 and HL60 cells were cross-linked with formaldehyde and processed according to our standard qChIP protocol. Messenger RNA expression analysis was performed by quantitative PCR with reverse transcription (RT–PCR). The PCR primers used for qChIP and mRNA analysis are listed in Supplementary Table 1. For RNA interference, HeLa cells were transfected using lipofectamine with siRNA oligonucleotides against PRMT1–6 or ASH2, and collected after 72 h. For complementary DNA overexpression, 293T cells were transfected using calcium phosphate with plasmids encoding Myc-tagged PRMT6 or haemagglutinin (HA)-tagged PRMT4 (both in wild-type or catalytically inactive dn mutant forms). Transfected cells were harvested after 48 h. For PRMT methyltransferase assays, Myc–PRMT6 or HA–PRMT4 were immunoprecipitated from transfected cell lysates and incubated with synthetic H3 tail peptides as substrates, the peptides were run in an SDS–PAGE gel, and H3R2 methylation was revealed by immunoblotting with the anti-H3R2me2a antibody. For H3K4 methylation assays, we used a purified TAP-ASH2/MLL complex. Tandem affinity purification (TAP) was performed following published protocols. Incubation of the complex with H3 tail peptides was performed in the presence of ^{3}H-Adomet. The peptides were run in an SDS–PAGE gel and incorporation of the radioactive methyl group was revealed by autoradiography. For peptide pull-down assays, U937 nuclear extracts were incubated with synthetic histone H3 tail peptides bearing the indicated methylation marks. All experimental procedures are described in detail in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 April; accepted 10 August 2007.
Published online 26 September 2007.

1. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
2. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–719 (2007).
3. Guccione, E. *et al.* Myc-binding-site recognition in the human genome is determined by chromatin context. *Nature Cell Biol.* **8,** 764–770 (2006).
4. Tenney, K. & Shilatifard, A. A. COMPASS in the voyage of defining the role of trithorax/MLL-containing complexes: linking leukemogensis to covalent modifications of chromatin. *J. Cell. Biochem.* **95,** 429–436 (2005).
5. Wysocka, J., Myers, M. P., Laherty, C. D., Eisenman, R. N. & Herr, W. Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3–K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. *Genes Dev.* **17,** 896–911 (2003).
6. Hughes, C. M. *et al.* Menin associates with a trithorax family histone methyltransferase complex and with the *hoxc8* locus. *Mol. Cell* **13,** 587–597 (2004).
7. Steward, M. M. *et al.* Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nature Struct. Mol. Biol.* **13,** 852–854 (2006).
8. Hampsey, M. & Reinberg, D. Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation. *Cell* **113,** 429–432 (2003).
9. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120,** 169–181 (2005).
10. Kouskouti, A. & Talianidis, I. Histone modifications defining active genes persist after transcriptional and mitotic inactivation. *EMBO J.* **24,** 347–357 (2005).
11. Bedford, M. T. & Richard, S. Arginine methylation: an emerging regulator of protein function. *Mol. Cell* **18,** 263–272 (2005).
12. Krause, C. D. *et al.* Protein arginine methyltransferases: evolution and assessment of their pharmacological and therapeutic potential. *Pharmacol. Ther.* **113,** 50–87 (2007).
13. Maurer-Stroh, S. *et al.* The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo, PWWP and MBT domains. *Trends Biochem. Sci.* **28,** 69–74 (2003).
14. Sims, R. J. III & Reinberg, D. Histone H3 Lys 4 methylation: caught in a bind? *Genes Dev.* **20,** 2779–2786 (2006).
15. Strahl, B. D. *et al.* Methylation of histone H4 at arginine 3 occurs *in vivo* and is mediated by the nuclear receptor coactivator PRMT1. *Curr. Biol.* **11,** 996–1000 (2001).
16. Huang, S., Litt, M. & Felsenfeld, G. Methylation of histone H4 by arginine methyltransferase PRMT1 is essential *in vivo* for many subsequent histone modifications. *Genes Dev.* **19,** 1885–1893 (2005).
17. Daujat, S. *et al.* Crosstalk between CARM1 methylation and CBP acetylation on histone H3. *Curr. Biol.* **12,** 2090–2097 (2002).
18. Bauer, U. M., Daujat, S., Nielsen, S. J., Nightingale, K. & Kouzarides, T. Methylation at arginine 17 of histone H3 is linked to gene activation. *EMBO Rep.* **3,** 39–44 (2002).
19. Covic, M. *et al.* Arginine methyltransferase CARM1 is a promoter-specific regulator of NF-κB-dependent gene expression. *EMBO J.* **24,** 85–96 (2005).
20. Schurter, B. T. *et al.* Methylation of histone H3 by coactivator-associated arginine methyltransferase 1. *Biochemistry* **40,** 5747–5756 (2001).
21. Torres-Padilla, M. E., Parfitt, D. E., Kouzarides, T. & Zernicka-Goetz, M. Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature* **445,** 214–218 (2007).
22. Wysocka, J. *et al.* WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell* **121,** 859–872 (2005).
23. Couture, J. F., Collazo, E. & Trievel, R. C. Molecular recognition of histone H3 by the WD40 protein WDR5. *Nat. Struct. Mol. Biol.* **13,** 698–703 (2006).
24. Han, Z. *et al.* Structural basis for the specific recognition of methylated histone H3 lysine 4 by the WD-40 protein WDR5. *Mol. Cell* **22,** 137–144 (2006).

25. Ruthenburg, A. J. *et al.* Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nature Struct. Mol. Biol.* **13,** 704–712 (2006).
26. Schuetz, A. *et al.* Structural basis for molecular recognition and presentation of histone H3 by WDR5. *EMBO J.* **25,** 4245–4252 (2006).
27. Dou, Y. *et al.* Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nature Struct. Mol. Biol.* **13,** 713–719 (2006).
28. Kirmizis, A. *et al.* Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation. *Nature* doi:10.1038/nature06160 (this issue).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Natoli, G. I. Dellino and all the members of our group for discussions; T. Kouzarides for PRMT4/CARM1 expression vectors and for communicating unpublished data; S. Richard, M. Wainberg and C. Invernizzi for PRMT6 vectors; D. Reinberg and P. Trojer for the Flag–ASH2 vector; and P. G. Pelicci for continuous support. This work was supported by a fellowship from the Italian Federation for Cancer Research (FIRC) to E.G., by grants from the Italian Association for Cancer Research (AIRC) to B.A. and by a grant from the Deutsche Forschungsgemeinschaft (DFG) to B.L.

Author Contributions E.G. and B.A. conceived the work and designed the experiments. B.A. supervised the project and wrote the manuscript. E.G. and F.M. performed ChIP, E.G. and C.B. performed biochemical experiments, and F.C. constructed expression vectors. M.C. performed the statistical analysis of ChIP data. H.S. and B.L. purified the TAP–ASH2 complex.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.A. (bruno.amati@ifom-ieo-campus.it).

METHODS

Antibodies. The antibodies used for ChIP were the following: from Bethyl Laboratory against menin (BL342), from AbCam against H3R2me2a (ab8046), H3K36me3 (ab9050), H3K4me3 (ab8580), total H3 (Ab1791), H3R26me2 (ab8047) and PRMT6 (ab47244), from Upstate against H3K27me3 (07-449), from Santa Cruz against the myc tag (9E10, sc-40) and the His tag (sc-8036). Rabbits were immunized with KLH-coupled peptides from human WDR5 (LENDKTIKLWKSDC) or ASH2 (CHVETEVDGRRSPPWEP) and affinity-purified sera were used for ChIP.

Cell culture, mRNA analysis and chromatin immunoprecipitation. P493 and HL60 cells were grown in RPMI medium/10% FCS supplemented with non-essential amino-acids and penicillin/streptomycin. Total RNA was prepared with the RNeasy kit (Qiagen). Complementary DNA was produced using the reverse-transcriptase SuperscriptII (Invitrogen). Quantitative chromatin immunoprecipitation (qChIP) was performed as described[3,29]. Real-time PCR reactions on ChIP-derived genomic DNA or cDNA were performed with SYBR Green Reaction Mix (Applied Biosystems), or with primer/probe combinations based on the Universal ProbeLibrary Set (Roche). PCR primer sequences are provided in Supplementary Table 1. For P493 cells, this work is based in part on the re-analysis of qChIP data and mRNA expression values measured in our previous study[3]. This applies to qChIP on histone marks and menin, but only on promoter sites, and thus excludes all qChIP data on non-promoter sequences, as well as on WDR5 and ASH2. All data in P493 cells were produced in the presence of tetracyline, that is in a condition in which the exogenous c-*myc* transgene was repressed[3,30].

Statistical analysis of ChIP data. Statistical analysis was as previously described[3]. Spearman rho and *P* values for the (counter-) correlations discussed are indicated in the Figures.

Cloning and protein expression. For expression as His-tagged proteins in transfected 293T cells, sequences encoding the chromo-domains of CHD1 and HP1β were PCR-amplified from cDNA and cloned with the Gateway Technology (Invitrogen) into pDEST 26. For *in vitro* translation of GST–WDR5 with the TnT T7 Coupled Reticulocytes Lysate System (Promega) the construct was subcloned into pDEST 15.

Peptide pull-down assays. Histone H3 tail peptides were purchased from Global Peptide. Peptides were resuspended in PBS at 2 mg ml^{-1}. The sequence of the unmodified peptide was NH2-ARTKQTARKSTGGKAP-GGY-[K(Biotin)]-OH. Approximately 10^9 cells were used per pull-down. Covalent modifications (R2me2a, K4me3, K4me2 or the double modification R2me2a/K4me3) were included during synthesis. For the preparation of nuclear lysates, U937 or 293T nuclei were resuspended in lysis buffer (50 mM Tris-HCl, pH 7.5, 1 mM EDTA, 1 mM EGTA, 0.5 mM Na_3VO_4, 10 mM Na β-glycerophosphate, 50 mM Na fluoride, 5 mM Na pyrophosphate, 1% Triton X-100 and the Roche protease inhibitor cocktail), briefly sonicated, precleared for 1 h with avidin beads and incubated with beads conjugated with 10 μg of each peptide. The beads were washed 5 times with washing buffer (50 mM Tris-HCl, pH 7.5, NaCl 150 mM, Triton X-100 0.2%).

Peptide methylation assays. Myc-tagged PRMT6, catalytically inactive PRMT6dn (bearing the triple amino acid substitution VLD to KLA)[31], HA-tagged PRMT4 and catalytically inactive PRMT4 were transfected into 293T cells. Following immunoprecipitation with anti-Myc (9E10) and anti-HA antibodies, the recovered PRMTs were incubated with 2 μg of histone H3 tail peptides for 2 h at 37 °C in HMT buffer (20 mM Tris-HCl, pH 8.0, 4 mM EDTA, 1 mM PMSF, 0.5 mM DTT) with 100 mM Adomet (Sigma-Aldrich). Reactions were stopped by addition of ×5 SDS–PAGE loading buffer. Following SDS–PAGE the specific di-methylation on Arg 2 was detected using AbCam anti-H3R2me2a (ab8046).

Given that ASH2 is a common component of SET1/MLL-family complexes[4–7] and that Flag-tagged ASH2 is associated with a K4 tri-methyltransferase activity[32], we stably expressed a TAP-tagged ASH2 protein in HEK293 cells. TAP purification was performed essentially as previously described[33]. The purified TAP–ASH2 complex is analogous to previously characterized SET1/MLL complexes and contains the catalytic subunit MLL4 (SwissProt ID Q9UMN6): its detailed analysis will be described elsewhere (H.S. and B.L., manuscript in preparation). For peptide methylation assays, 2 μg of recombinant peptides were incubated with 400 ng of the TAP–Ash2 complex in HMT buffer with 1.5 μl ^{3}H-Adomet (15 Ci mM^{-1}; Amersham) for 1 h at 30 °C. Reactions were stopped by addition of ×5 SDS–PAGE loading buffer, followed by SDS–PAGE and autoradiography. The same assay was repeated with Flag–ASH2 immunoprecipitates in place of the TAP–ASH2 complex as a source of methyltransferase activity[32], with similar results.

29. Frank, S. R., Schroeder, M., Fernandez, P., Taubert, S. & Amati, B. Binding of c-Myc to chromatin mediates mitogen-induced acetylation of histone H4 and gene activation. *Genes Dev.* **15,** 2069–2082 (2001).
30. Pajic, A. *et al.* Cell cycle activation by c-myc in a burkitt lymphoma model cell line. *Int. J. Cancer* **87,** 787–793 (2000).
31. Xie, B., Invernizzi, C. F., Richard, S. & Wainberg, M. A. Arginine methylation of the human immunodeficiency virus type 1 Tat protein by PRMT6 negatively affects Tat interactions with both cyclin T1 and the Tat transactivation region. *J. Virol.* **81,** 4226–4234 (2007).
32. Pavri, R. *et al.* Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II. *Cell* **125,** 703–717 (2006).
33. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17,** 1030–1032 (1999).

naturejobs

JOBS OF THE WEEK

Every year, the announcement of who has won a Nobel prize provokes debate as scientists reappraise the discoveries, and the significance of the work is reaffirmed to the general public. But one aspect of the awards tends to garner less media coverage: the geography of the prizes — in particular, the nationalities and geographical locations of the winners.

In fact, this year's physics prize is something of an anomaly. It is the first time in eight years that none of the winners has been either American or based in the United States. A glance at the winners of the three scientific Nobel prizes of the past 10 years shows just how unusual that result is. For the years 1998–2007, nearly 60% of the winners were working in the United States at the time they received their prize, although roughly 30% of them were not American. Other winners were working in Australia, Denmark, France, Germany, Israel, the Netherlands, Japan, Russia, Switzerland and the United Kingdom.

The statistics are similar if Nobel success is gauged by institution across the history of the prize. In the top ten universities, ranked by the number of laureates, only two are outside the United States —and both of those are in Britain. Pole position goes to Harvard, which has 21 laureates to its name.

Of course, such a cursory analysis can highlight only where leading scientists end up —and the statistics reflect, in part, the locations of large, well-funded universities, as well as political strife and the quest for the best education.

The big question is whether the increasing globalization of science will see a change in geographical distribution of Nobels. Japan, for example, stated in 2000 that it aims to secure 30 Nobel prizes by 2050 (see *Nature* **413,** 560–564; 2001). At the same time, talented young researchers are not now automatically heading for the United States to further their education and career. True, recent competitiveness legislation in the United States is aimed at reasserting the nation's scientific dominance, but nevertheless, maybe we will soon be seeing Nobel prizes being won by American expatriates — rather than the other way round.

Gene Russo, acting editor of *Naturejobs*

CONTACTS
Acting Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young
(+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com

France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Southwest UK/RoW:
Nils Moeller (4953)
Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
North Germany/The Netherlands:
Reya Silao (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:**
Tom Hancock

***Naturejobs* online production:**
Jasmine Myer
US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
Chiyoda Building,
2–37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe
Tel: +81-3-3267-8765
e-mail: a.watanabe@natureasia.com

DNA masters

As genetic testing becomes more common, the need rises for experts to interpret the results. **Ricki Lewis** reports.

P. SALOUTOS/CORBIS

As the fruits of the Human Genome Project continue to make headlines, a related field has been quietly gestating — genetic counselling. Its practitioners are hybrid professionals, combining expertise in human genetics with the ability to communicate their knowledge to patients and families. "This profession is very much like a small, quaint community that is well known among some circles, but is on the verge of being discovered by the rest of the world," says Janey Youngblom, associate director for the master's degree programe in genetic counselling at California State University, Stanislaus in Turlock.

With the recent expansion of genetic-disease screening programmes in newborns, the improved safety of amniocentesis for prenatal diagnosis and a slew of direct-to-consumer genetic-testing websites, genetic counsellors are more in demand than ever.

The term 'genetic counselling' was coined by geneticist Sheldon Reed in 1947, referring to the advice he provided to physicians whose patients had inherited diseases. The inaugural master's degree programme in genetic counselling — at Sarah Lawrence College in Bronxville, New York — saw its first class graduate in 1971, and this course has served as the model for the current 32 programmes in the United States.

The field is still small — the National Society of Genetic Counselors (NSGC) in Chicago, Illinois, has just 2,300 members, says Caroline Lieber, director of the graduate programme at Sarah Lawrence. In Europe, the first genetic-counselling programme started in 1992, from the University of Manchester, UK, and 14 other nations are playing catch-up, with the most new efforts in Japan.

At the inaugural meeting of the Transnational Alliance for Genetic Counselling in Manchester in May 2006, educators representing 45 programmes and 15 professional organizations, from 18 nations, met to discuss training. There, the term was consolidated. Jacquie Greenberg, an associate professor in human genetics at the University of Cape Town in South Africa, summarizes it: "Genetic counselling helps people understand how genetic disease affects their lives."

Evolving roles

When the field began, patients weren't genetics-savvy. Today, the Internet has changed that picture dramatically — but experts are needed to explain the science and to take psychosocial factors into account.

"People often form initial perceptions based on what they read on the web, and it can be difficult to make them see something from a different point of view," says Siobhan Dolan, associate professor of reproductive genetics at the Albert Einstein College of Medicine of Yeshiva University in New York. A nutrigenetics website, for example, might present population-level information on the association between a gene variant and a disease before studies have been done on the validity of extrapolating it to individuals. Likewise, patients using direct-to-consumer genetic-testing services might confuse high risk with general population risk, notes Dolan. Generally, high risk for, say, the *BRCA* gene and breast cancer, is suspected for patients who have a young first-degree relative with breast cancer.

"This profession is well known among some circles, but is on the verge of being discovered by the rest of the world."
— Janey Youngblom

C. EDMONDSON

The counsellor's role transcends the provision of information and correction of misinformation. "We help patients make sense of the options, and provide reassurance and support," says Jennifer Fitzpatrick, director of the genetic-counselling programme at McGill University in Montreal, Canada. For example, if a prenatal screen shows a high risk of Down syndrome,

the parents-to-be could very well think about termination. "Our role is to do damage control and put the information in perspective," she says. In this common scenario, the counsellor explains that a positive result from the screening indicates elevated risk, not a diagnosis, and explains further testing options.

"People often form initial perceptions based on what they read on the web."
— Siobhan Dolan

J. COATE/MARCH OF DIMES

The genetic counsellor is an advocate for the family, says Karen Marder, professor of neurology at Columbia University in New York. As a physician, Marder works closely with a genetic counsellor, who navigates families through the labyrinth of choices that accompany testing for Huntington's disease. "I can't imagine a multidisciplinary team without one," she says.

A genetic counsellor is especially helpful in rare diseases with which the physician is not familiar. Helga Toriello, director of genetics services at Spectrum Health in Grand Rapids, Michigan, gives the example of spinal muscular atrophy. There's no need to sequence the gene to diagnose it, she notes. "Don't test for all the mutations for a disease; do the common ones first."

Perhaps the counsellor's greatest contribution is time. In an hour-plus session or in multiple meetings, a counsellor discusses the effects of mutations and how they are transmitted, sketches a pedigree, talks to relatives about risks, tests and treatments and explores feelings such as fear, anxiety and guilt.

And the field is growing fast. "The profession began with paediatric and prenatal care, but now includes cancer, cardiovascular disease, neurology and ophthalmology," says Beverly Yashar, director of the genetic-counselling graduate programme at the University of Michigan in Ann Arbor. A newer turf is public policy. "Policy-makers, like the public, have genetic terms in their lexicon but are not always sure what they mean," says Luba Djurdjinovic, executive director of the non-profit Ferre Institute in Binghamton, New York, which provides community-based genetics services. The institute's annual DNA Day lecture informs the state's legislature.

"The profession now includes cancer, cardiovascular disease, neurology and ophthalmology."
— Beverly Yashar

M. VLOET

Training and traits

A genetic counsellor combines technical knowledge with characteristics such as empathy, curiosity, and comfort with new technology and with people who have disabilities, says Youngblom.

The career trajectory begins with an undergraduate degree in a biological or social science. Master's degree programmes in genetic counselling include coursework in all areas of genetics, as well as statistics, psychology, and 50–200 supervised cases. Certification in the United States requires passing an exam given by the American Board of Genetic Counseling in Olathe, Kansas. Licensing — a step beyond certification — is progressing on a state-by-state basis. Genetic counsellors are certified by the Health Professionals Council in South Africa and by the National Health Service in Britain.

"Don't test for all the mutations for a disease; do the common ones first."
— Helga Toriello

M. TABER-LIND

Many genetic counsellors start their careers in science or medicine, but desire more time with patients, says Yashar, who earned her PhD in yeast genetics and then her master's in genetic counselling. The preponderance of females in the profession reflects the initial focus on prenatal care, the history of female domination of counselling fields and the flexibility of many jobs, which eases childcare concerns, says Youngblom.

Varied workplaces

Most genetic counsellors work in clinical settings. According to the NSGC's 2006 workplace survey, 38% of its members work in university medical centres, 31% in hospitals, 8% in diagnostic laboratories and 5% in private practices. The rest work at health-maintenance organizations, universities and in biotech, pharma or Internet companies.

Although the day-to-day tasks may be similar in the different settings, the degree of patient involvement varies. At Columbia University, says Marder, a genetic counsellor develops long-term relationships with families. In the corporate sphere, the job may be less personalized and broader. A genetic counsellor for Myriad Genetics in Salt Lake City, for example, interprets the results of hereditary cancer tests, keeps up with the literature, prepares educational material and presents training programmes.

Research is also possible. At Myriad, counsellors help improve breast-cancer tests. In universities, they can work with families enrolled in clinical trials, says Yashar. Opportunities are also opening up in the direct-to-consumer market. For example, people who seek testing for some mutations online can take a tissue sample at home and send it to a lab. The results are then sent to the patient and a genetic counsellor, with whom the patient can discuss the results — imperative for disorders in which not all people who inherit the mutation develop the disease.

The future

As genetic discoveries proliferate, more genetic counsellors will be needed to bridge the gap between what consumers think they know and what scientists have learned.

Prenatal care is at the forefront of the integration of genetic counselling into medical practice. Some obstetric practices, for example, routinely test for cystic fibrosis carriers, with nurses or physicians explaining common findings, says Yvette Conley, assistant professor of health promotion and development at the University of Pittsburgh in Pennsylvania. Patients with rare allele combinations are referred to genetic counsellors. "We will be educating all kinds of health-care professionals," says Lieber.

Djurdjinovic predicts that the marriage of new information and outreach may even create a new type of practitioner. "We can expect to see a new wave of genetic counsellors practising in small communities to support the application of new genetic tests for common conditions," she says. As microarray technologies become more common, Djurdjinovic sees genetics professionals becoming more important to the appropriate application of testing and interpretation of test results. "Maybe a new professional will emerge: the genomic counsellor," she says. ■

Ricki Lewis is the author of *Human Genetics: Concepts and Applications*.

J. SOHM; VISIONS OF AMERICA/CORBIS

Golden opportunities

With a variety of federal positions and a fledgling life-sciences sector, northern Virginia offers plenty of opportunities, from bench researcher to programme officer. **Ted Agres** reports.

Just west of Washington DC, where federal funding and science projects are lost and found in the halls and budget proposals of the US Congress, lies northern Virginia. Given its proximity to the nation's capital, it's hardly surprising that the area has numerous government jobs in science and science policy, ranging from biodefence to geology, economics and social sciences. Indeed, the government agencies of northern Virginia help decide how to divvy up the science budgets of Capitol Hill.

As you drive into northern Virginia, the District of Columbia's massive concrete buildings and imposing security barriers give way to the businesses and pricy housing of Arlington, Alexandria and Falls Church. Some two million people call northern Virginia home. Loudon County, one of its four main jurisdictions (along with Arlington, Fairfax and Prince William) is the nation's fastest-growing large county.

The northern Virginia city closest to Washington is Arlington, home to several science agencies. Chief among these is the National Science Foundation (NSF), with 800 to 900 science- and engineering-related employees. About 450 of these are programme directors, overseeing specific research areas as well as the awarding and administration of grants. NSF programme directors generally keep up with new developments and emerging trends and stay in touch with leading scientists.

Unlike their counterparts at the National Institutes of Health, programme directors at the NSF have the authority to override peer-review panels' grant funding recommendations. Because of this, NSF programme directors must be particularly well-respected in their fields. "It's extremely important for them to have that credibility, especially among those whose grant proposals get denied," says Joseph Burt, the NSF's director of human resources.

P. G. SPYROPOLUS

Carol Bessel can continue her independent research while working as a programme director at the National Science Foundation.

About 250 programme directors are 'temporary rotators', on assignment from universities and non-profit organizations for 1 to 3 years. About 100 to 150 of these jobs are available during the course of a year. A doctorate and at least 6 years of academic or non-profit research experience are needed. But unlike many federal jobs, US citizenship is generally not required. Rotators are paid their academic salaries, but permanent programme directors average $140,000 annually — not bad, even in a region with a very high cost of living. The recently enacted America competitiveness legislation would allow a substantial increase to the NSF's budget during the next 3 years. If this happens, the number of programme directors at the agency might well increase.

Looking from a different angle

Being a programme director affords scientists the chance to see their fields from a unique perspective. "You might think it's just administrative work, but every idea that will be published in a journal in 5 years is coming across your desk in the form of a proposal or an idea," says Alan Tessier, a programme director in the division of environmental biology.

Programme directors can also continue their own research. Under the NSF's independent research and development programme, they are encouraged to spend a day a week on their own research at a lab not funded by the foundation. Carol Bessel, a full-time programme director in the NSF's inorganic, bioinorganic and organometallic chemistry division, continues her work with carbon nanofibres at the Naval Research Laboratory across the river in DC. "You get to see the science and you can also participate in it," says Bessel.

Down the street from the NSF is the Defense Advanced Research Projects Agency (DARPA), the Pentagon's secretive 'high risk/high reward' think-tank. Housed in a pleasant yet non-descript glass office building, DARPA seeks 25–30 programme managers each year for 4–6-year assignments, welcoming

scientists from industry and academia who have what the agency terms 'far-side' ideas.

"We do fundamental research without a specific application in mind," says spokeswoman Jan Walker. "But it has to have some military relevance." DARPA salaries are commensurate with experience, and US citizenship is required for security clearances.

Many DARPA-funded projects have civilian benefits — lasers, advanced computing, the global positioning system and Internet technologies all benefited from early DARPA support. Mathematician Benjamin Mann is running seven projects in DARPA's defence sciences office, including applying algebraic topology to massive data sets and statistics. In addition to the military, the work may have applications to neuroscience and biology, he says. He and other programme managers draft solicitations for external grants and contracts, participate in the award selection process (along with Pentagon brass), and oversee the research being conducted at universities and institutes. Although nearly everything DARPA does with its sponsored research is classified, DARPA-funded scientists are encouraged to publish their findings.

K. SIMPSON

The Defense Advanced Research Projects Agency employs scientists from industry and academia to do basic research with military relevance.

Dig this

Conducting research of an entirely different sort, the US Geological Survey (USGS) headquarters is nestled in a series of office parks in Reston, Virginia, 20 minutes northwest of Arlington. The USGS acquires and assesses data involving geography, geology, hydrology and biology to help the government minimize the effects of natural disasters and manage natural resources. Its large, eight-pointed star-shaped complex houses about 600 scientists who coordinate and oversee research conducted at USGS science centres in every state, including a regional seismic monitoring network for earthquakes and volcanoes. Some scientists at the headquarters oversee the conduct of competitive grants to universities and research institutions; others assess energy supply resources for public policy decision-making. Still others manage mapping activities, including a large geospatial imaging programme.

"This region is just burgeoning in technical activities."
— John Mathieson

"We do a tremendous amount of hiring for biologists, geologists and hydrologists," says Roxanne Tipton, chief of the staffing and classification office for the USGS eastern region. Typically 15–25 doctorate-level science positions are open in Reston, she says, and scientists with supervisory skills and and project-management experience can earn from $110,000 to $143,000 per year. Citizenship is required for senior-level positions.

Federal money also means plenty of jobs for scientists at private contractors and consultancies in northern Virgina (see 'Worlds apart'). Non-profit think-tanks SRI International and the RAND Corporation in Arlington, for example, offer science policy and other research employment. "This region is just burgeoning in technical activities," says John Mathieson, director of the Center for Science, Technology, and Economic Development at SRI. "It's a very dynamic place for jobs for scientists." ■

Ted Agres is a freelance writer based in Laurel, Maryland.

WORLDS APART

Although separated physically only by the Potomac River, northern Virginia and suburban Maryland are worlds apart when it comes to biotechnology. Nearly 200 of Maryland's 370 biotech companies have opened shop in Montgomery County, north of Washington DC. Even promoters of northern Virginia admit that their side of the river pales by comparison, hosting only about 60 of the state's 175 biotech firms. But some see northern Virginia as brimming with biotech potential.

Maryland's biotechs include some of the nation's oldest and most prominent, including MedImmune, Celera and Human Genome Sciences. Northern Virginia, by contrast, is a relative newcomer. It acquired its first biotech, American Type Culture Collection, in 1998. A non-profit organization, the firm occupies lab space in the Innovation @ Prince William Technology Business Park, a 600-hectare business and technology park anchored by the life-science campus at George Mason University in Prince William county.

In addition to its proximity to government facilities such as the National Institutes of Health (NIH) and the Food and Drug Administration, Maryland's biotech strength also derives from its major research universities, including the University of Maryland in College Park and Johns Hopkins University in Baltimore. Johns Hopkins was the nation's leading NIH grant recipient in 2005, with 1,299 awards totalling $607 million. That year, George Mason University — the second-largest university in Virginia with nearly 30,000 students on three campuses — had only 17 NIH awards, at $29 million.

But together, government and private grants hit $85 million last year at George Mason University, says provost Peter Stearns, and funding has been growing at around 14% a year for the past several years. Two years ago the university started its Center for Applied Proteomics and Molecular Medicine. And it broke ground last month on a $42-million biocontainment laboratory and bioscience research building on 4 hectares adjacent to the university's Prince William county campus.

Supporters of northern Virginia biotech industry hope that the Janelia Farm Research Campus — the Howard Hughes Medical Institute's $500-million research complex in Ashburn — will be a magnet for biotech development when fully operational in 2009 (see *Nature* **443,** 128–129; 2006). "We're already seeing some interest from outside companies, but it's too early to say what will happen," says Michelle Snyder, communications director for the Northern Virginia Technology Council. Janelia has recruited about half of the 450 people it is aiming to employ, according to director Gerry Rubin. "We are still looking for everything from group leaders to administrative assistants," he says.

Oblivious to turf differences, outsiders view northern Virginia, suburban Maryland and the District of Columbia as one region — the nation's sixth-largest biotech cluster, according to a 2004 study by the Milken Institute.

"If you talk to people from overseas, they don't care about state lines," says Terry Woodworth, life-sciences director at the non-profit Virginia Center for Innovative Technology in Herndon. "It's one area." **T.A.**

NATURE|Vol 449|18 October 2007

MOVERS

Bob Watson, chief scientific adviser, UK Department for Environment, Food and Rural Affairs

2005–07: Director, International Assessment of Agricultural Science and Technology for Development, Washington DC
2001–07: Chief scientist and senior adviser for sustainable development, World Bank, Washington DC
1996–2001: Scientific adviser, then director, environment department, World Bank

Trained as a chemist at Queen Mary College, London, Bob Watson had no intention of entering the public-policy arena. He researched how halogen atoms, such as chlorine, interact with ozone to form chlorine monoxide radicals.

During a postdoc at the University of California, Berkeley, Watson saw that chemistry had social relevance as he watched mentor Harold Johnston debate with the Nixon administration over the impact of supersonic transport on stratospheric ozone depletion. When chlorofluorocarbons were found to trigger that depletion, Watson's expertise was suddenly in high demand. "Careers are made as much on luck as judgement," he says.

Joining NASA as a scientist in the Jet Propulsion Lab in Pasadena, Watson left as director of the science division. During his time there, he was asked to direct a national assessment of ozone depletion. Watson cited seven other recent assessments and sought not to duplicate efforts, but build international consensus. "Policy-makers need a single scientific assessment by the world's best scientists," he says.

His next move was to the White House, into the president's Office of Science and Technology Policy. He continued to lead assessments, co-chairing a working group of the Intergovernmental Panel on Climate Change (IPCC) and the Global Biodiversity Assessment. In 1996, he began a decade at the World Bank, directing its environment department, and then becoming the bank's chief scientist, while also serving as chair of the third IPCC report.

Former World Bank colleague Ian Johnson, now chairman of IdeaCarbon, a UK-based carbon market analysis firm, says Watson was one of the rare scientists who saw science not as outside of public policy, but as integral to it — especially as a driver of development change. "Scientists don't often like to see consensus and compromise, but Bob understands the sometimes painstakingly slow need to listen and share information to reach consensus," says Johnson.

In his latest move, Watson has accepted three positions: as the chief scientific adviser of the UK Department of Environment, Food and Rural Affairs; as a professor at the University of East Anglia; and as director for strategic development of a unique collaboration of UK academics at the university's Tyndall Centre for Climate Change. Although he has no plans to participate in yet another international assessment, he doesn't dismiss the idea. "My entire career has been a random walk, so one never knows," he says. ■

Virginia Gewin

NETWORKS & SUPPORT

A question of balance

In a meta-analysis of peer-review procedures during grant applications, Lutz Bornmann at the Swiss Federal Institute of Technology in Zurich, Switzerland, and his colleagues discovered a bias against women (see *Nature* **445,** 566; 2007).

Our team at the Swedish Research Council has now studied all 17,000 grant applications received by the council during 2003–05. We found that in Sweden there is little evidence of gender bias: success rates for men and women were, in most cases, roughly the same (see www.vr.se).

But closer inspection showed that there were some discrepancies: women had less success with fellowships to be postdocs abroad, as well as in long-term grants for prominent research environments and in nearly all types of grant in the field of medicine.

Why should this be? The cause doesn't seem to be fewer women applying nor because there were fewer women in the peer-review groups or in high-ranking positions. The most likely explanation is 'career age' — the number of years that have passed since applicants earned their PhDs. Overall, success rates for both sexes were higher for increasing career age. But on average, women applying for project grants had a lower career age than men, which skewed the balance. This explained the grant discrepancies but it accounted for only half the difference in the medical applications. (Career age is irrelevant to postdoctoral fellowships.)

Another factor that deserves attention is the relative quality of the applications. Bibliometric methods offer an indicator of quality and can compare the scientific output of a large group of men with that of a large group of women. We studied 225 applications for postdoctoral fellowships, but found no bibliometric differences between the sexes.

Our data cannot explain the remaining discrepancies in medicine. But other studies point to similar problems. The US RAND Corporation uncovered obstacles for women in medicine (S. D. Hosek *et al. Gender differences in major federal external grant programs*; RAND, 2005). And women had less success than men in a recent European Research Council call for life-science grants (see http://tinyurl.com/326jxj). This suggests that it would be worthwhile to carry out an international comparative study.

For its part, the Swedish Research Council now plans to monitor closely the decision-making process to look for explanations for any discrepancies. ■

Gunnel Gustafsson, Carl Jacobsson and Carolyn Glynn are at the Swedish Research Council.

POSTDOC JOURNAL

Simulating life

It never ceases to amaze me how one can be a biomedical scientist without ever getting under a fume hood or pipetting into a test tube. During my doctoral research, I wrote computer programs that simulated the electrical activity of the heart, providing insights into heart function that experiments would be hard pressed to supply. Now, for my postdoctoral research, I'm still sitting at a computer. This time I'm constructing mathematical models that simulate the metabolic processes regulating the composition of the human body, particularly in infants. All without dissecting a heart or cradling a newborn in my arms.

Although it may seem otherwise, models aren't simply figments of an overactive imagination. The best are based on real data and reproduce phenomena that have already been observed, while predicting other phenomena that have not. In another sense, however, models are imaginary. They are born as sketches on a piece of paper, mature into a bundle of ones and zeros inside a computer, and retire as text in a journal article or book. No one has ever seen a model under a microscope or felt a model's heart beat.

My research is possible only because others poke and prod living organisms. But given the choice, I'd rather be on the simulation end. I may not be dealing with the 'real' thing, but it's astounding what can still be discovered. ■

Peter Jordan is a visiting fellow at the National Institute of Diabetes and Digestive and Kidney Diseases in Bethesda, Maryland.

NATURE|Vol 449|18 October 2007

Red

The gift of life.

Melissa Yuan-Innes

From the operating table, Rosa Clarke watched the anaesthetist insert a line into her right hand. For a 15-year-old, she must have had difficult veins, because he had to poke her twice before the burgundy blood flashed back up the catheter. She didn't flinch.

Usually I don't enter the OR until the residents have prepped and draped the patient, but this one was special.

She was the youngest spontaneous donor we'd ever had, and arguably one of the smartest. I crossed to her side. "Are you sure?"

She nodded, but whispered, "My grandmother would kill me." She looked about as pale and thin as the hospital sheet she was lying on. Her long, wavy, cinnamon-coloured hair was the brightest thing in the OR.

When she first walked into my clinic, all I saw was the hair. Recessive gene jackpot. I scanned her face: another check mark for mild to moderate freckles. Then she'd made me see her as an individual. First of all, she pointed out the high miscarriage rate. Even though we did our best at immunohistogenetic matching, our success rate was worse than any other transplant surgery.

I'd started to believe my own hype. The applause of my colleagues at conferences was addictive enough without the gifts, donations and documentaries of reci-moms. Just that morning, I had downloaded a mini-movie in which the reci-mom held a Doptone to her very pregnant belly, cut to a PG-13 shot of the labour, and ended with her holding newborn baby girl. "Thank you, Doctor Fletcher, for helping our dream come true."

But Rosa had said: "Isn't it true that more than half the fetuses die? Even with you doing the surgery? And you're supposed to be the best, aren't you?"

Then she refused to answer questions about her parents. "I left that blank on purpose. You can ask everything about me, but not my parents."

I'd objected to this question myself. We don't know how much of intelligence is genetic, and careers don't equal intelligence. The board psychiatrist had intoned, "IQ tests aren't practical. So for most spon donors, parental occupations are the best we'll get. Birth mothers want smart babies."

So this strange, strong, red-haired girl was a woman after my own heart. At least, she was until she refused to answer my questions about her partner. "Rosa, it would help your embryo to answer these questions. The ones who don't get placed, well, you'd have to go back to the old-fashioned choices of abortion, adoption or raising the baby yourself." Her green eyes revealed nothing. "Tell me anything. Like, does he have red hair?"

Rosa half-snorted. "No."

Too bad. I wrote, non-red hair. "Anything else?"

"Anything to sell my baby?" she snapped.

I studied her for a minute. "Anything to play the game, so your baby has a better chance at life."

"Ha." She blinked away a tear. "What happened to just wanting a baby?"

I shook my head and didn't answer.

Her fists clenched. "I want it to have a good home. That's why I…" She cut herself off and stared at me. "I want to screen the birth parents back. I want to see them before they get…implanted."

"Done." Highly unusual, but no one would have to know.

"Now, about — him." She thought a little. "He plays the trombone. He's good at science, and he has a great French accent." A smile flickered. "He — likes children."

In the silence, I said softly, "That's fine, Rosa. Thank you."

She signed the consent forms so hard that her pen ripped the page.

So now, with Rosa before me, I asked her one more time. "Are you sure you want to donate your embryo?"

She turned her head away. "Just do it."

The residents stopped talking. I felt the anaesthetist's and nurses' eyes on me. Finally, I said, "All right."

Once she was under and prepped and draped, I made an incision just above the pubic hair, to open the abdominal layers down through the peritoneum. A resident retracted the bowels. I gently incised the uterus with a robotic scalpel. Under the microscope's guidance, I removed an embryo the size of my pinky nail as well as its placenta. The nurses sighed with relief when I slipped it into a nutri-gro bottle.

The anaesthetist woke up Rosa and wheeled her to recovery, and after a quick room-clean, the reci-mom rolled in, praying. Her husband stroked her hand. "Don't worry, Lorna. We'll have our little redhead yet, God willing."

JACEY

The embryo would have one recessive, red-haired gene from its mother. Artificial downregulation of the melanocortin-1 receptor could guarantee red hair, but it was still unpopular, especially with the religious candidates.

Two hours later, I ducked into step-down recovery to tell Rosa all had gone well. "Someone can take you home."

She swung her legs over the side and grabbed the bed for balance. "I'll call a cab."

"Rosa, you're still…"

"I'm fine." She started to stand, then froze. I followed her gaze to a young man in the hallway. "Booker!"

He caught her before she fell. "Baby."

"Booker." He stroked her hair, and she wrapped her arms around him tightly. "Booker, you came."

They looked so young, so right together. My throat ached. His dark cocoa fingers moved tenderly against her red, red hair.

A cry strangled in my throat.

"Doctor!" Rosa held herself just out of his arms. "Doctor, Booker is not the father." He opened his mouth, but she sliced it off with a look.

I'd listed the father's race as unknown when the reci-parents signed. Now I had possible information that one of my patients wanted suppressed, and the other, more powerful patient, would want known.

There was no test for race — there were more molecular genetic similarities between races than differences. In the end, I'd probably have to do a second spontaneous donation. The chances of finding another compatible donor were very small. The chances of a couple wanting a half-white, half-black baby twice-removed were even smaller. The chances of the embryo surviving another transplant were almost nil.

Inside my office, I leaned against the door and wept. ■

Melissa Yuan-Innes is an emergency physician who has sold fiction to L. Ron Hubbard's *Writers of the Future*, *Weird Tales*, *Tesseracts7*, *Island Dreams* and *Open Space*, as well as to various small presses.